ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

PhD THESIS

Chipo MUFUDZA

MIXTURE MODEL BASED DATA CLUSTERING

DEPARTMENT OF STATISTICS

ADANA-2017

ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

MIXTURE MODEL BASED DATA CLUSTERING

Chipo MUFUDZA

PhD Thesis
DEPARTMENT OF STATISTICS

We certify that the thesis titled above was reviewed and approved for the award of degree of Doctor of Philosophy by the board on .....

.....	.....	.....
Prof.Dr. Güzin YÜKSEL	Prof.Dr. Hamza EROL	Prof.Dr. Sadullah SAKALLIOĞLU
SUPERVISOR	MEMBER	MEMBER
.....	.....	
Doç Dr. Turgay İBRİKÇİ	Doç Dr. Tayfun SERVİ	
MEMBER	MEMBER	

This PhD Thesis is written at the Department of Statistics of Çukurova University.
Registration Number:

Prof. Dr. Mustafa GÖK
Institute Director

Note: The usage of presented specific declarations, tables, figures, and photographs either in this or in any other reference without citation is subject to "The Law of Arts and Intellectual Products" number 5846 of Turkish Republic

ABSTRACT

PhD THESIS

MIXTURE MODEL BASED DATA CLUSTERING

Chipo MUFUDZA

ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF STATISTICS

Supervisor : Prof.Dr. Güzin YÜKSEL
Year: 2017, Pages: 136
Jury : Prof.Dr. Hamza EROL
: Prof.Dr. Sadullah SAKALLIOĞLU
: Assoc. Prof.Dr. Turgay İBRİKÇİ
: Assoc. Prof.Dr. Tayfun SERVİ

The study focused on exploring the use of mixture models for data clustering techniques on data with complex heterogeneous properties. Both parametric and nonparametric mixture models were here considered. In particular, Poisson mixture regression models and Zero Inflated Poisson mixture models were applied and analysed for the parametric case as mixtures of Gaussian kernel densities were considered for the nonparametric case. Although, parametric mixtures assume clusters and components to be the same, these are observed different by either using graph theory techniques via voronoi diagrams or infinite componential models in the nonparametric case. Determination of better clustering method using mixture models was enabled through classification via clustering analysis. Results shows that although number of clusters from the parametric mixtures can be the same with nonparametric, these approaches are generally very different. Nonparametric mixtures produce better clusters as evidenced by their ability to improve classifier performance and hence compacting heterogeneity in a more efficiently way whilst parametric are a bit rigid. The ability by mixture models to allow inferences on cluster properties via densities and parameters over traditional clustering methods makes them a better clustering methods. We conclude that although mixture models are better than hierarchical models, nonparametric mixture should be preferred over parametric as they are more robust, flexible and produce better clusters.

Key Words: Poisson Mixture models, Zero Inflated Poisson Mixture, BIC, Gaus sian kernel density function, Density Based Silhouette information.

ÖZ

DOKTORA TEZİ

KARMA MODEL TABANLI KÜMELEME

Chipo MUFUDZA

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İSTATİSTİK ANABİLİM DALI

Danışman : Prof.Dr. Güzin YÜKSEL
Yıl: 2017, Sayfa 136
Juri : Prof.Dr. Hamza EROL
: Prof.Dr. Sadullah SAKALLIOĞLU
: Doç Dr. Turgay İBRİKÇİ
: Doç Dr. Tayfun SERVİ

Bu çalışma, bir popülasyondaki karmaşık heterojenlik özelliklerini araştıran kümeleme tekniklerine dayalı karma modellerin incelenmesi üze-rinde yoğunlaşmıştır. Hem parametrik hem de parametrik olmayan karma modeller farklı durumlarda incelenmiştir. Özellikle Poisson karma regresyon modelleri ve Sıfır şişirilmiş Poisson karma modelleri parametrik durum için incelenirken Gauss çekirdek yoğunlukları da parametrik olmayan durum için incelenmiştir. Parametrik karmalar, kümeleri ve bileşenleri aynı varsayarken, bunlar parametrik olmayan durum için voronoi diagramları ve sonsuz bileşen modelleri aracılığı ile grafik teori teknikleri kullanılarak farklı şekilde gözlemlendi. Karma modeller kullanılan daha iyi kümeleme yöntemleri kümeleme analizi yolu ile sınıflandırılarak belirlendi. Sonuçlara göre, parametrik ve parametrik olmayan karmalardaki kümelerin sayıları aynı olmalarına rağmen bu yaklaşımlar genellikle çok farklıdır. Parametrik olmayan karışılar, sınıflandırıcı performansını arttrıma ve dolayısıyla heterojeliği daha verimli bir şekilde sıkıştırma yenetenekler ile kanılandığı üzere daha iyi kümeler üretir. Bu nedenle, karma modellere, geleneksel kümeleme yönetemleri üzerindeki yoğunluklar ve parametreler yoluyla küme özelliklerinde çkarımlara izin verme yeteneği onları daha iyi kümeleme yönetemleri haline getirir. Parametrik olmayan karma daha sağlam, esnek ve daha iyi kümeler ürettikleri için parametrik olarak tercih edilmelidir.

Anahtar Kelimeler: Poisson Karışımı modelleri, Sıfır şişirilmiş Poisson karma, BIC, Gauss çekirdek yoğunluk fonksiyonu, Yoğunluk tabanlı siluet bilgisi.

GENİŞLETİLMİŞ ÖZET

Tez, model tabanlı kümeleme yoluyla popülasyonlarda gözlenmeyen heterojenite ile ilgilenmeye odaklanmıştır. Gözlemlenmeyen heterojenlik birkaç yönden ele alınabilir ancak, biz hem parametrik hem de parametrik olmayan yaklaşımlar için karma modellerinin kullanımı üzerinde yoğunlaştık. Bunlar homojenliğe alternatif bir görünümdür ve verinin kendi parametre değerlerinin kümeleri ya da sırasıyla Everitt ve Hand (1981) ve Magidson ve Vermut (2002) tarafından formüle edilen ve üstü kapalı bahsedilen popülasyonun gözlenemeyen özelliklerini bulunduran karma popülasyonlardan geldiği varsayılır. Bu karma modeller daha sonra çeşitliliği yakalar ve verilen popülasyon içindeki kümeler veya gruplar vasıtasıyla heterojenite oluşur.

Parametrik karma modeller McLachlan ve Peel (2000) ve Fraley ve Raftery (1998) tarafından yapılan çalışmalar temel alınarak geliştirilmiş ve bir analiz bölüm 3 de verilmiştir. Parametrik olmayan karma modeller, bölüm 4 de incelenmiştir ve bunlar Azzalini ve Toreni (2007) tarafından yapılan çalışmalara dayanarak ve Benaglia ve ark. (2009b) tarafından yapılan analizler dahil olmak üzere geliştirilmiştir. Bu nedenle, heterojenlikle baş edebilmek için etkili varsayımlarla homojenliği varsayan geleneksel yollara kıyasla heterojen popülasyonda verilen farklı gruplardaki farklı parametre değerleri ve karmaların dağılımları aracılığı ile heterojenliği düzenleyebilen karma modellerin kullanımı üze-rinde yoğunlaşacağız. Bu çalışmanın amacı, bu karma modellerin karmaşık heterojenlik özelliklerini açıklamada, farklı dağılım ve paramet-relerin nasıl yardımcı olabileceğini bir analiz ile göstermektir. Odak noktası ayrıca, parametrik ve parametrik olmayan modeller arasındaki bu heterojen zorlukları aşmak için en iyi yöntemin belirlenmesini de içerir. Bu, bölüm 5 'te sunulan sınıflandırma algoritmaları yoluyla karma modellerdeki etiketler üzerinde karşılaştırmalı analizlerle elde edildi.

Karmaşık heterojen özelliklere sahip olduğu için, özellikle bu tezdeki tüm örnekler için kalp hastalığı verileri kullanılmıştır. Veriler, Poisson

hata tipi ile belirtilen karma sayım verileridir. Verilerde, bölüm 3'de verilen genelleştirilmiş karma Poisson modeli analizinde görüldüğü gibi ekstra dağılım özellikleri de gözlenmiştir. Bu nedenle, Poisson karma modelleri, heterojenlik için parametrik karma modelleri altında geliştirilmiş ve analiz edilmiştir. Özellikle, Genel Poisson karma modelleri ve Poisson karma regresyon modelleri farklı parametre değerleri ile Grün ve Leish(2007) tarafından incelenmiştir. Bunlara aynı zamanda değişken değişkenli modeller de dahil olmak üzere çok değişkenli ve modüler değişkenlik gösteren modeller dahil edilmiştir. Sıfır frekans verilerini de karşılayan karma modeller, 3. bölümde Sıfır şişirilmiş Poisson karma modelleri şeklinde tanıtıldı. Parametre tahminleri, McLachlan ve ark. (1997) tarafından açıklandığı gibi EM algoritması ile yapıldı.

Her ne kadar parametrik karmalar heterojeniteye işaret etse de, tüm karmalar için aynı şekil ve bileşen olarak aynı küme varsayılmaktadır. Daha sağlamlık ve esneklik eklemek ve heterojenliğin sıkılaştırılma düzeyini artırmak için, kalp hastalığı verisi için varsayımsız modeller geliştirildi ve parametrik olmayan karma modeller şeklinde analiz edildi. Bunlar sadece farklı bir yaklaşım değil, yoğunluk fonksiyonunun tahmininde ve bölüm 4'de araştırılan popülasyondaki gruplamayı belirlemede daha esnek bir yol bulundurmaktadır. Parametrik olmayan karma modelleri iki şekilde gösterilir: Delaunay grafikleri ile birleştirilen grafik modüllerinden şekillenen kümelerin bulunduğu Gauss fonksiyonlu çekirdek yoğunluklarının karmalarından ya da Azzalini ve Toreni (2007) tarafından açıklanan şekliyle voronoi diyagramındaki çiftlerden. Bu, R'deki PdfCluster paketi kullanılarak analiz edildi ve veri formatını değiştirmek için küme paketi kullanıldı.

Çekirdek türünün karma modelin şekli üzerinde herhangi bir etkisinin olmadığı, ancak bant genişliği pürüzsüzleştirme parametresindeki değişimin karma modelde belirgin etkilere yol açtığı sonucundan sonra yoğunluk tahmin edicileri seçildi. Bant genişlikleri, Scott ve Sain (2005) tarafından verilen önerilere uygun olarak kullanıldı. İkinci yöntem, Be-

naglia ve ark. (2009b) tarafından belirtildiği gibi EM benzeri algoritmayı kullanarak tahmin edilen parametreler ile bileşenler arasında koşullu bağımsızlığı varsayan sonsuz bileşen modelini kullandı. Bu R'deki mvnp fonksiyonu aracılığı ile bir mixtools paketinden sağlandı.

Tezin son kısmında parametrik ve parametrik olmayan karma modelleri karşılaştıran bir analiz üzerinde duruluyor. Amaç, karma modellemeden üretilen grup etiketleri olduğu için daha iyi etiketleri belirlemekti. Bu etiketli verileri farklı yöntemlerle elde etmek için Weka 8.0'daki farklı sınıflandırıcılar kullanıldı. Herhangi bir sınıflandırıma problemi için doğruluk geliştirilebilirse, etiketlerin kalitesi daha iyi olarak kabul edileceğinden, daha iyi bir yöntem olduğu kararına varılmıştır. Etiket bilgileri doğrudan etiketlenmemiş verilerden çekiliyor ise doğrulukta iyileşme sağlanabilir sonucu Sakai ve ark. (2016) tarafından verilen bir sonuçtur.

Sonuçlar, parametrik karmalar kullanıldığında parametrik olmayan karma modellere kıyasla heterojenliğin farklı şekilde düzenlendiğini göstermektedir. Parametrik karma modelleri temsil etmek için poisson karma regresyon modelleri kullanılmıştır. Eşzamanlı ve eşzamanlı olmayan değişken modeller altında analiz edilmiştir. Her iki durumda da, 2 bileşenli model kalp hastalığı için karma dağılımı en iyi temsil eden modeldir. Bu farklı bileşenler, kalp hastalığına değişen oranlar ve risklerle farklı şekilde görüldü. Bu bileşenlerin kalp hastalıklarının gelişimi için yüksek risk (kalp hastalığınada hızlı ilerleme için) ve düşük riskli (yavaş ilerleme) grupları olduğunu varsaydık. Poisson karma regresyon modelinde iki model de 2 bileşen modunda sonuçlanmasına rağmen, birlikte gelen model en düşük Bayes Bilgi Kriteri ($BIC = 762.06$) ile eşlik etmeyen değer (835.29) e göre daha iyi bir model olduğunu kanıtladı. Tablo (3.7.) da gösterildiği gibi iki farklı bileşendeki farklı parametre tahmin değerleri de birbirinden farklılık gösterir. Sıfır şişirilmiş Poisson karmaı regresyon modeli (verilerde sıfır değerli yüksek frekansın varlığı göz önüne alınır) ile yapılan daha ileri analizler, bu model kullanılarak verilerin sıfır frekansının yüksek olmasında dolayı modeli ve esnekliği geliştirir. Bu, diğer tüm modeller arasında en düşük BIC

değeri 731.17 ile gösterildi. Modelde 0.3686 olasılık ile hastalığın gelişme ihtimali bulunmayan ya da gelişme olasılığı düşük olan (sıfır dejenere olmuş) ve eşlik eden değişken model ($BIC = 731.17$) kullanarak 0.632 olasılıkla hastalığı gelişme olasılığı yüksek olanlar olmak üzere iki grup bulunmaktadır. Böylece parametrik karma modeller kalp hastalığı verilerinde kullanıldığında bireylerin yüksek risk ve düşük risk olmak üzere iki kümede toplandığı sonucuna varılabilir.

Parametrik olmayan karma model sonuçları, Gauss çekirdeği yoğunluk tahmin edicileri ile gözlemlerin kümelenmesinde grafiksel yöntemler kullanıldığını gösterdiği için daha ilginçtir. Bu sonuçlar parametrik karmalarda aynı bileşenli küme varsayımının aksine sayısının değişmesine neden olan farklı sayıda bileşenler ile sonuçla-nabilir. Kümelerin kalitesini belirlemek için yoğunluk temelli siluet (dbs) bilgilerini kullandık ve böylece en iyi modelleri belirledik. Genel olarak çekirdek işlevi için uyarlanabilir bant genişliği kullanıldığında, sabit bant genişliği modellerine kıyasla daha yoğun ve iyi ayrılmış kümeler elde edildiğini gözlemledik. Bu, kümelerin çakışmadığı ve gözlemlerin doğru sınıflandırıldığına dair bir göstergenin, bölünmüş, negatif değerli dbs çizelgelerinin bulunmamasıyla kanıtlandı.Yine kalp hastalığı verileri kullanıldığında, sabit bant genişliği dbs kriterine göre 2 bileşen ve 3 kümeli iken adaptif bant genişliği ile 2 bileşen ve 3 kümenin en iyi model olduğu görülür. EM benzeri algoritma ile çözülen bileşen modeli, en düşük ortalama hata kareler değerine sahip olduğu için en iyi model olarak 2 bileşenli 3 blok modelini ortaya çıkarmıştır. Parametrik olmayan karma modeller farklı yöntemler ile kullanıldığında heterojenliği yakalamada daha iyi daha gelişmiş sonuçlar vereceğinden kümelerin daha kaliteli olacağı sonucuna varabiliriz. Farklı bileşenlerin değişen kümeleri olmasına rağmen, kalp hastalığı verisi için en iyi modeller parametrik olmayan karma modellerdeki bileşen ve küme sayılarının parametrik karma modellerdeki sayılar ile aynı olabileceğini gösterir.

Ayrıca parametrik olmayan ve parametrik karma modeller arası-

ndaki daha ileri analizler, daha iyi bir kümeleme tekniği ve dolayısıyla daha iyi veri etiketleri belirlemek için yürütülmüştür. Farklı yöntemlerle etiketlendirilen verilere farklı sınıflandırma algoritmaları uygulanmıştır. Sonuçlar, parametrik olmayan karmalar, verilerin etiketlenmesi için kullanıldığında, sınıflandırıma performansının artmasıyla sınıflandırma doğruluğunun arttığı gösterilir. Böylece, parametrik olmayan karmalarnın parametrik karma modelleri üzerinde daha iyi etiketler verdiğinden daha iyi model olduğunu çıkarabiliriz. Parametrik karmaın sağlamlığının yanı sıra yorumlanmasının ve keşfedilmesinin de kolay olduğu belirtilmelidir.

Bu nedenle, karma modellerin model tabanlı kümeleme yaklaşımında kullanılması ve uygulanması durumunda, adaptif bant genişliği modelleri ile parametrik olmayan karmanın, sabit modeller üzerinde ve aşırı parametrik karma modelleri üzerinde tercih edilmesi gerektiği sonucuna varabiliriz. Parametrik olmayan karma modeller parametrelerin üzerinde daha iyi etiketler ve kümeler sağlar, böylece daha sonra kolayca yorumlanabilir. Heterojenlik, bu nedenle, veri türüne bağlı olarak karma modeller aracılığıyla çeşitlendirilmiş bir şekilde ele alınabilir. Parametrik karma modeller, heterojen özellikleri yakalamak için kolaylıkla formüle edilirken ve analiz edilebilirken, parametrik olmayanlar karmaşıktır, ancak daha iyi sınıflandırmalara neden olduğundan daha sağlamdırlar. Karma modeller, sadece veri kümelemesi yapmayıp, yoğunluk fonksiyonları aracığı ile her bir kümenin özelliklerinin ne anlama geldiğini yorumlayıp açikladıkları için veri kümelemesinde bütüncül yaklaşım sağlarlar. Parametrik olamayan karmalar ileride Dbs kriteri aracılığı ile küme içi ve kümeler arası uzaklıkları sağlar. Bu sebep-le karma model tabanlı kümelemeler daha iyi kümeleme yönetemleridir.

Yayınlanmış Makaleler

Mufudza C., Erol H., 2016. Poisson Mixture Regression Models for Heart Disease Prediction. Computational and Mathematical Methods in Medicine, Article ID 4083089, doi:10.1155/2016/4083089. **(SCI), Appendix 6.6 bakınız**

İncelenen Makaleler

Mufudza C., Erol H., 2017. Heart Disease Diagnosis Using Nonparametric Mixture Models. FILOMAT. **(SCI)**

Konferanslar

Mufudza C., Erol H., 2017. Heart Disease Diagnosis Using Nonparametric Mixture Models. 2nd International Conference On Advances In Natural and Applied Sciences, ICANAS 2017, Antalya, TURKEY April 18-21, 2017. **Presenter**: Chipo Mufudza, **Appendix 6.6 bakınız**

ACKNOWLEDGEMENTS

My sincere gratitude goes to a number of people for the different insights, input and knowledge during the course of the research. Firstly to my Prof.Dr. ***Güzin YÜKSEL***, co-supervisor ***Prof.Dr. Hamza EROL*** and the whole supervisory committee for their invaluable knowledge, expertise and encouragement, without which it was going to be impossible to complete the doctorate.

The positive attitude, prayers and jovial face that I would always receive from my lovely husband ***Ndarama Zidana*** was amazing. You endured all the loneliness, bad meals and worst moments in my absence as a commitment to see me finish. Your special words always "Chiuya Chips mudiwa, but usatya nguva irikutofamba zvakutoita (Come Chips dear, but do not worry time is moving its going to be fine)" were like a shot in the arm. Indeed golden gifts like you are given only of the Lord. Thank you darling.

I wish to express my heartfelt thanks to my mom too, ***sissy T*** as we call her, her unconditional moral and spiritual support was paramount. I am grateful to all my siblings and friends who have become sisters like Vongai Mudombo. Ronald Musizvingoza your input cannot be forgotten thank you. The Lord bless you all. The Turkish government scholarship was also there with the financial support allowing me a chance to explore and release my potential, I am grateful.

Finally, to the ***Lord God Almighty***, great are you Lord and faithful in **ALL** your doings. If it was not of your mercies and sufficiency I would have succumbed, **Ebenezer** (This far the Lord has taken me)!!!!!

CONTENTS

1. INTRODUCTION 1
   1.1. Purpose and Importance of Study 1
   1.2. Research questions 3
   1.3. Proposed Methodology And Material 3
2. LITERATURE REVIEW 7
   2.1. Introduction 7
   2.2. Heterogeneity 7
   2.3. Finite Mixture Models 8
      2.3.1. Mixtures of Univariate Data 10
      2.3.2. Heterogeneity and Mixture Models 11
      2.3.3. Mixture Model Based Clustering 15
   2.4. Parameter Estimation 18
      2.4.1. Maximum Likelihood Estimators(MLE) 18
      2.4.2. EM Algorithm 19
      2.4.3. Question of Optimality 21
   2.5. Density Based Nonparametric Clustering 23
3. FINITE MIXTURE MODELS 27
   3.1. Introduction 27
   3.2. Finite Mixture Regression Models 29
      3.2.1. The Poisson Mixture Regression Model 31
      3.2.2. EM for GLM Finite Mixture Models 33
   3.3. Application: Heart disease prediction 34
      3.3.1. Heart Disease and Classification 35
      3.3.2. Data 36
      3.3.3. Dependent Variable Distributions 37
   3.4. Generalised Poisson Regression Model (PR) 39
      3.4.1. Why GLM Poisson 39
      3.4.2. Residual Plots 42

3.5. Why Mixture Models on Heart Disease? 43
3.5.1. EM clustering in Weka 44
3.6. Application: Poisson Mixture Regression (FMPR) 45
3.6.1. FMPR without Covariates 45
3.6.2. Two Component Covariate Poisson Mixture Model 46
3.6.3. FMPR with Covariates 47
3.6.4. Properties for Two Component FMPR Model 48
3.6.5. Parameter Estimates 50
3.7. Zero Inflated Poisson Mixture Model 52
3.7.1. Two Component ZIPM Model Properties 52
3.7.2. Cluster Distributions 53
3.7.3. Parameters 54
3.8. Summary 55
4. NONPARAMETRIC MIXTURE MODELS 57
4.1. Introduction 57
4.2. Nonparametric Density Estimations 57
4.2.1. Kernel Density Estimation (KDE) 58
4.2.2. Kernels 59
4.2.3. Kernel Density Estimations Selection 60
4.3. Clustering via Nonparametric Mixtures 61
4.3.1. Nonparametric Mixture Model Identification 62
4.3.2. EM for Multivariate Nonparametric Mixtures 63
4.4. Nonparametric Mixtures via Graph Theory 65
4.4.1. Graph Theory Procedure 67
4.4.2. Semiparametric Mixture Modeling 68
4.4.3. Density Based Silhouette Information (DBS) 69
4.4.4. Bandwidth Selection 71
4.4.5. Motivation 72
4.5. Nonparametric Mixture Models Applications 73
4.5.1. Data 73
4.5.2. Packages Used 73

4.6. Using KDE with Adaptive Bandwidth 74
4.6.1. Cluster Scatter Plots under Adaptive Bandwidth 77
4.6.2. Cluster Summary Properties under Adaptive Bandwidth 78
4.6.3. Two Component Two Cluster Model Properties 79
4.6.4. Surface Densities for the 2-2 Model 79
4.7. Nonparametric Mixtures with Fixed Bandwidth 80
4.7.1. Cluster Scatter Plots for Fixed Bandwidth Models 83
4.7.2. Summary: Fixed Bandwidth Models 83
4.7.3. Fixed Bandwidth Best Model Properties 84
4.8. Conditionally Independent Densities for Nonparametric mixtures: MvnpEM 85
4.8.1. Model Results 86
4.8.2. Component Integrated Error Densities 87
4.8.3. Clusters from MvnpEM 88
4.8.4. Componential Cumulative Distribution functions 89
5. DATA LABELING VIA MIXTURE MODELS 91
5.1. Introduction 91
5.2. Classification 92
5.2.1. Classifiers 93
5.2.2. Naive Bayes 93
5.2.3. Logistic 94
5.2.4. IBK 94
5.2.5. Random Forest 94
5.3. Classification via Nonparametric Mixture Models 95
5.3.1. Nonparametric Mixture Labels Evaluation 96
5.3.2. Classifier Performance and Output 98
5.4. Classification via Parametric Mixture Models 98
5.4.1. Parametric Mixture Models Label Evaluation 99
5.5. Nonparametric versus Parametric labeling 101
5.6. Summary 102

6. RESULTS AND CONCLUSIONS 103
6.1. Introduction 103
6.2. Use of Finite Mixture Models 103
6.2.1. Use of Finite Mixture Regression Models 104
6.3. Use of Nonparametric Mixture Models 105
6.4. Data Labeling via Mixture Models 105
6.5. Heart Disease Risks 106
6.6. Further Research Recommendations 108
APPENDIX 123

LIST OF FIGURES

3.1. Dependent variable density: a) The density of num grouped by sex, b) density of num grouped by chest pain, c) density of num grouped by thal, d) density of num grouped by slope, e) density of num grouped by fbs, f) density of num grouped by restecg. . . 37

3.2. Dependent variable density: a) The density of num grouped by ca, b) density of cholestrol grouped by num, c) density of thalcd grouped by num, d) density of Age grouped by num, e) density of trestbp grouped by num, f) density of oldpeak grouped by num. 38

3.3. (a) Pearson deviance for the Poisson Regression (b) Residual deviance for the Poisson regression. 41

3.4. Residual Poisson regression Plots: a) Residual plot, b) QQ plot, c)Deviation and d) Leverage 42

3.5. Rootogram for FMPR Model without Covariates 47

3.6. The FMPR Model with Covariates: a) FMPR Models with concomitant variable and b) FMPR Models without concomitant. 48

3.7. 2 Component FMPR Model with Covariates: a) Rootogram for concomitant variable model b)Parameter Confidence intervals concomitant model, c) Rootogram for model concomitant variable and d) Parameter Confidence Interval for Model without concomitant. 49

3.8. ZIP Model: a) Rootogram for Concomitant variable Model, b) Rootogram Model without concomitant, c) Confidence interval for parameters, and d) Model Selection graphs. 54

4.1. Kernels Approximations: a) Different Kernel functions used, b) Different bandwidth for Gaussian kernel 60

4.2. Voronoi Diagram: The left plot displays the Voronoi tessellation (dashed lines) for the points and superimposed Delaunay triangulation (solid lines). In the right plot some edges have been removed from the Delaunay triangulation, those connecting pairs of observations with density below the threshold c represented by the plane in the right plot 68
4.3. Dbs diagrams for different p principal components: a) $p = 2$, b) $p = 3$ 75
4.4. Dbs using adaptive bandwidth for different p principal components: a) $p = 4$, b) $p = 5$. 76
4.5. Dbs diagrams for different p principal components: a) $p = 8$, b) $p = 10$. 76
4.6. Nonparametric cluster scatter plots with 2 components 77
4.7. Nonparametric Surface Densities 80
4.8. Variable Surface densities for clustered data: a) Age, b) sex, c) cholesterol, d) resting blood pressure, e) maximum heart rate, f) resting electrocardiographic, g) chest pain, h) defects 81
4.9. Dbs diagrams for fixed bandwidth and different p principal components: a) $p = 3$, b) $p = 4$ 82
4.10. Dbs diagrams for fixed bandwidth and different p principal components: a) $p = 3$, b) $p = 4$ 84
4.11. Integrated square error for densities f_{ij} component $i = (1,2)$ and block j $j = 1,2,3$. 88
4.12. Component Distribution 3 block Model with: a) two components, b) 3 components, c) 4 components and d) 5 components . 89
4.13. Cumulative function for Components:a) 2 component model, b) 3 component model, c) 5 component model, d) 7 Component model 90
0.1. Densities for the variables: a) Age, b) cholesterol, c) resting electrocardiographic, d) resting blood pressure, e) Maximum heart rate, f) ST depression 129

0.2. Densities for the variables: a) sex, b) Chest pain, c)Fasting blood sugar, d) Exercise induced angina, e) slope of peak exercise, f) Number of major vessels 130
0.3. Densities for the variables: a) number of diagnosis, b) Defects effects 131
0.4. Variable Surface densities for clustered data: a) number of vessels, b) Number of diagnosis , c) exercise induced angina d) slope of exercise and e) depression induced by exercise 133

LIST OF TABLES

3.1. Poisson regression Model Output 40
3.2. Chi-square test on parameters 41
3.3. The PR model and Simple PR summary 43
3.4. EM Results in Weka 45
3.5. Poisson Mixture without Covariates 46
3.6. FMPR with Covariate Models 48
3.7. Summary: FMPR Models 50
3.8. Parameter estimates 51
3.9. ZIP Models 53
3.10. Summary of the ZIPR Models 53
3.11. Parameter Values of the ZIPR Models 54
4.1. Kernel weights 59
4.2. Cluster Summary Under Adaptive Bandwidth 78
4.3. 2-2 Model DBS Cluster Summary 79
4.4. Cluster Summary for Models with Fixed Bandwidth 85
4.5. Best Model properties with Fixed Bandwidth 85
4.6. Integrated Square Errors 87
5.1. Naive Bayes as test base 97
5.2. IBk as test base 97
5.3. Nonparametric Labeling Classifier Results 98
5.4. Naive Bayes as test base 99
5.5. IBK as test base 99
5.6. Random forest as base 100
5.7. Parametric Labeling Classifier Results 100
5.8. Naive bayes as test base 101
5.9. Random forest as test base 101
0.1. Cleveland Heart Disease Variables 127

1. INTRODUCTION

1.1. Purpose and Importance of Study

The complexity involved in many real life scenarios due to either observed or unobserved heterogeneity has remained a mouthful task in data modeling and analysis for centuries. Although there have been efforts to deal with such data complexities via data transformations, real problems normally have complex heterogeneous characteristics which can never be totally outdone by simple data transformations. It should therefore, never be forgotten that real world processes probably contains both elements of groupings and heterogeneity as such the purpose on any model must be to capture the main features of the process. Frequently, just data transformations fails to capture all features of a real problem leaving heterogeneity and groupings in data analysis a paradox.

These heterogeneous characteristics are found in almost every life discipline including climatology, rainfall data, ecology, medicine, economy, marketing segmentation and energy data to mention but a few. Statistical methodologies which helps deal with clustering and heterogeneity still remain a challenge in all these real life disciplines. Simple data transformations have been in use for centuries in simple problems like regression, discriminant analysis etc, to cab heterogeneity, they seem not to have totally eradicated the paradox especially with huge data being produced in the 21st century and beyond. In some cases researchers have managed to deal with heterogeneity making clustering complex or vise versa. Examples include the work by Kedem et al. (1990), Shimizu (1993) and recent work by Suhaila et al. (2011), Suhaila et al. (2013) who addressed heterogeneity in the environmental factors by using mixed multinomial distributions in rainfall occurrence and amount. However, clustering the locations and defining proper cluster attributes of these mixed distributions has not been considered as robust methodologies maybe required to factor in the unobserved

heterogeneity.

It is the aim of this work therefore to come up with best ways to cater for clustering complexities in the presence of heterogeneity in big data scenarios. In many medical researches nonparametric methods have been used to cater for heterogeneity, however, this diverges from proper inferences about the population groups in the case of data clustering. These challenges has lead to the birth and use of model based clustering using mixture distributions. Although these mixture model based clustering techniques have been in existence for over half a century, they have been only in active use over the past decade, McNicholas (2011). They can be applied in many disciplines which exhibit heterogeneous tendencies in their characteristics. Examples of how mixture models can be used to cater for heterogeneity and hence clustering include medical, life insurances, climatology and many more as in Fraley and Raftery (1998), McLachlan and Peel (2000), McNicholas and Murphy (2010), Ng et al. (2006), Schlattmann (2009), Erisöǵlu et al. (2011), Boswell et al. (1979). Thus they are used in the different real life processes especially medical biology since biological and medical data normally display notorious heterogeneity of responses due to stimuli and treatment.

Heterogeneity, whether observed or unobserved is not however, a non unique feature since it can never be uniquely done away with and hence the need to check for ways to improve the problem. Mixture models are of late being used in market segmentation since its an area which has vast heterogeneous problems due to different customer characteristics, Greene (2005). Thus we focus on the use and application of mixture model based clustering on different choice models with unobserved heterogeneity in big data. The hypothesis to deal with heterogeneity in big data can be either unknown or known but segmented in a single variable in the case of multivariate data or both.

1.2. Research questions

- How best can unobserved heterogeneity in different natural scenarios be addressed without doing away with the natural features of the population? Is a single distribution enough?
- Is assuming variable independence a reality even in big data?
- How do we identify clusters for correlated variables/observations ?
- How do we factor variable segmentation for multivariate big data when clustering?
- How best can we observe or infer cluster properties after cluster identification?
- Uniqueness: Does the identified clusters match the optimal clusters and why ?
- Does model based clustering helps over the traditional hierarchical clustering methods, if yes, in what sense?

1.3. Proposed Methodology And Material

This work will involve addressing latent heterogeneity in various disciplines and clustering homogeneous characteristics. As such appropriate mixture models will be chosen, which can be either continuous or discrete depending on the data type. Many application examples to be considered here will focus on big data although small data sets can also be considered. Thus a combination of mixed and mixture models may be considered which is basically dealing with variability, heterogeneity and clustering in a goal. Thus in general we wish to solve the problem of clustering via a model based approach using different mixture models of the form given by equation (1.1).

$$f(\mathbf{x}|\theta) = \sum_{i=1}^{k} \pi_i f_i(\mathbf{x}; \phi_\mathbf{i}) \tag{1.1}$$

Where $f_i(x;\phi_\mathbf{i})$ is called the component density with scalar vector of parameters ϕ_i for some $k \in Z$ and $\pi_i \in [0,1]$, $\sum \pi_1 = 1$. The equation (1.1) is referred as the k component mixture density function. Different examples of hetero-

geneity and mixture model based clustering are going to be considered in this work resulting in the function $f(x|\theta)$ of (1.1) having a variety of generalised mixture forms which can be:

1. Same distribution family but different parameter values and orientation (Gaussian mixtures, gamma mixture, etc)

$$f(x|\Theta) = \sum_{i=1}^{g} \pi_i \left(g(x|\theta_i)\right) \tag{1.2}$$

The mixture in (1.2) are of the same type such that the function $g(x|\theta_i)$ can be either univariate or multivariate and of the same type like all Gaussian with different orientation hence θ_i is different.

2. Different distributions (eg. mixture of Weibull and Gamma distributions)

$$f(x|\Theta) = \pi_1 g(x|\theta) + \pi_2 r(x|\eta) \tag{1.3}$$

The equation (1.3) is an example where $g(x|\theta)$ can follow say a gamma distribution whilst $r(x|\eta)$ follows a different distribution say weibull, resulting in a mixture of different parametric distributions with different parameter values too.

3. A mixture of mixed distributions whether of the same type or differently

$$f(x|\Theta) = \sum_{i=1}^{g} \pi_i \Phi\left(x|\theta_i\right) \tag{1.4}$$

The form of $\Phi(x|\theta_i)$ in equation (1.4) can also be a mixture as shown in Suhaila et al. (2011) when they used mixed gamma, mixed log-normal or mixed Weibull as a rainfall distribution so the mixture model $f(x|\Theta)$ will become a mixture of these mixed distributions.

4. Flexibility and assumption free mixture models via the use of non-

parametric mixture models using the kernel density estimators.

$$f_\theta(\mathbf{x_i}) = \sum_{j=1}^{m} \lambda_j \left\{ \prod_{k=1}^{r} f_{jk}(x_{ik}) \right\} \tag{1.5}$$

If we let z is the latent random variable with probability mass function (PMF) : $\chi \to [0;1]$, where $\chi = \{z_1, z_2, ..., z_k\}$ for an integer K. $\mathbf{x} = (x_1; x_2; .....x_r)$ is be a vector of repeated measurements on an observable outcome variable x whose marginal probability density function (PDF) takes the form equation (1.5) and f_{jk} denotes the PDF of $\mathbf{x_i}$ conditional on $z = z_k$ and it is given as an estimation using kernel density estimators.

A common formulation for the multidimensional kernel estimator by Scott and Sain (2005) is given by the product kernel estimator:

$$\hat{f}(x) = \sum_{i=1}^{n} \frac{1}{nh_1.....h_d} \prod_{j=1}^{k} K\left(\frac{x^{(j)} - x_i^{(j)}}{h_j} \right) \tag{1.6}$$

where x^j is the jth component of x, the univariate kernel K is usually taken to be a non-negative function centered at zero and integrating to one, and a different smoothing parameter h_j is chosen for each component.

Parameter estimation and optimality in the number of components will be considered via the EM algorithm and different methods described above will be applied as it seem fit for each relevant data set as described by Wei and McNicholas (2015), Brochado and Martins (2005), McLachlan and Rathnayake (2014), Ng et al. (1999), respectively. Due to the complexity of the heterogeneity in the mixture models to be considered in this work, the renowned Gaussian mixture models may fail to work hence the need for more detailed method which also deal with optimisation of parameters. Mixture of mixed model may have to be considered in some data sets depending on the nature of the variables. We therefore propose the use of packages for mixture models like **Mixtools** by Benaglia et al. (2009b), **FlexMix** by Grün et al.

(2008) and **PdfCluster** by Azzalini et al. (2014) in R. Other packages to be considered may include **MixEst** toolbox in Matlab developed by Hossein and Mash'al Hossein et al. (2015) which allows the optimisation of the mixture models via an EM algorithm by defining any given particular function being optimised.

2. LITERATURE REVIEW

2.1. Introduction

In this chapter, a review of analysis on heterogeneous populations, the use and application of mixture models both parametric and nonparametric, and their use on heterogeneous populations and parameter estimations will be discussed.

2.2. Heterogeneity

A heterogeneous population involves differences in a certain pattern or characteristic being observed. These heterogeneous attributes can be either observed or unobserved (latent) with no information known about it by the analyst. Latent heterogeneity is the dispersion of factors known and relevant for decision makers, but unknown to the analyst. In many cases, data is frequently analyzed as if it was obtained from a single population (homogeneous), although it is often unlikely that all individuals in a sample have the same set of parameter values. Unknown characteristics in a population and environments often results in exhibiting notorious heterogeneity in the data. Thus heterogeneity can come as a result of differences in parameter values across groups and within groups. Heterogeneity can be observed within different groups, for it is realistic to assume that individuals of different groups obey different response processes and relationships between variables. Reality do not have independence among all observations. It is realistic to assume that individuals within a group share certain influencing factors and hence have correlated observations as shown in Muthen (1989).

In this study no independence or homogeneity assumptions are going to be made on heterogeneous populations, hence the need to come up with a way to properly cluster data from a heterogeneous population. Capturing heterogeneity is not always easy but can be done in many ways: One way is to explain it using a scale parameter. It has been observed that pa-

rameterising heterogeneity in choice model with demographic components tend to exhibit superior statistical properties over models with imposed homogeneity. Secondly, the use of random parameters in logit/probit models account for heterogeneity by allowing model parameter to vary over a characteristic. However, these methods are not well suited to explain the source of heterogeneity and the heterogeneous group characteristics. Thus there is needy for a method which can capture the heterogeneity and represent the topology, classification and integration of the information involved. One such promising method is the use of mixture models.

2.3. Finite Mixture Models

An alternative view to homogeneity is that data come from a mixture of populations with their own sets of parameter values. This relates to statistical modeling called finite mixture analysis as explained by Everitt and Hand (1981). Mixture Models are an extension of standard parametric models and a tool which captures the richness of real life phenomena. They deal with latent classes or unobserved groups unlike the traditional models used in regression, discriminant analysis and log-linear analysis contain parameters that describe only relationships between the observed variables. Finite Mixture Models (FMM) also known as Latent Class models (LCM) differ from these by including one or more discrete unobserved variables, Magidson and Vermut (2002).

Finite mixture modeling is typically associated with inferences on the model and its parameters whilst model based clustering provides a partition of the data into groups of homogeneous observations. They provide a natural and mathematical way to encode heterogeneity in a population since they are ideal for density estimation, clustering and classification problems. FMM where first introduced and discussed by Newcomb (1986) as model of outliers and more detailed by Pearson (1894) and up to date they are still receiving popularity in model based clustering and classification. The basic idea of mixture models is a sample drawn from a population where the

label of the population is missing, then a homogeneous distribution from standard parametric family is inadequate to assess the characteristics of the data, Dai (2006). This gives them a peculiar feature of label switching which is permuting the label component labels and they also do not rely on the traditional modeling assumptions which are often violated in practice (linear relationship, normal distribution, homogeneity). Hence, they are less subject to biases associated with data not conforming to model assumptions. Also, for improved cluster or segment description (and prediction), the relationship between the latent classes and external variables (covariates) can be assessed simultaneously with the identification of the classes (clusters, segments).

If population is sampled at random then probability model is of the form given by equation (1.1). i.e: $f(\mathbf{x};\theta) = \sum_{i=1}^{k} \pi_i f_i(\mathbf{x};\phi_\mathbf{i})$ where models parameters are explained as in chapter 1. The equation (1.1) is referred to as the k component mixture density function. In most literature the component densities are assumed to be from the same parametric family. A finite mixture distribution can also be defined as a weighted linear combination of a finite number of simple component distributions. The standard density of the parametric FMM can also involve multivariate data like:

$$f(\mathbf{x}|\pi_1,\ \pi_2,\ .....\pi_g,\ \theta_1,\ ....\theta_g) = \sum_{i=1}^{g} \pi_i g(\mathbf{x}|\theta_i) \tag{2.1}$$

where $\pi_i \in [0,1]$, $\sum \pi_i = 1$ and $g(\mathbf{x}|\theta_i)$ is the density of the multivariate random variable $\mathbf{x}$ with parameter θ_i as in model by McNicholas and Murphy (2010). Recently, more mixture models with different parametric distributions have been used like mixture of Poisson distributions used for document classification in the field of information retrieval, Li and Zha (2006). The parameter in the parametric distribution $g(x|\theta_i)$ in equation (2.1) can be either discrete or continuous. In most model based clustering literature however, the Gaussian mixture models are the most commonly used, McLachlan and Peel (2000), McNicholas and Murphy (2010), Raftery and Banfield (1993),

Fraley and Raftery (2012). The Gaussian mixture model is given as:

$$f(\mathbf{x}|\theta) = \sum_{i=1}^{g} \pi_i \Phi \left(x_j | \mu_i, \sum_i \right) \tag{2.2}$$

where

$$\Phi \left(x_j | \mu_i, \sum_i \right) = (2\pi)^{-\frac{p}{2}} \left| \sum_i \right|^{\frac{-1}{2}} \exp -\frac{1}{2} (x_j - \mu_i)^T \sum_i^{-1} (x_j - \mu_i)$$

is the density of the multivariate Gaussian random variable **x** with mean vector μ_i and covariance matrix $\sum_i (i = 1, ..., g)$ also used in Ng et al. (2006). Thus the vector θ of unknown parameters is given by

$$\theta = \left(\pi_1, \ ....\pi_g, \ \mu_1, \ .... \ \mu_g, \ \sum_1 \ ..... \sum_g \right).$$

FMM analysis yields a probabilistic clustering approach. This means that although each object is assumed to belong to one class or cluster, it is taken into account that there is uncertainty about object class membership. In human populations like epidemiological problems, survival and reliability data, count data with over-dispersion in many cases often results in model-based structures which identifies the groups for individuals or patients and labels them. Although the traditional K means clustering and hierarchical methods can be used, very little information can be inferred into the groupings which in many cases does not follow a homogeneity. Thus there is need to capture the heterogeneity through equation model (1.1), which allows finite combinations of distributions from different distinct parametric families.

2.3.1. Mixtures of Univariate Data

Latent heterogeneity in some univariate data have been also dealt with using mixture models. Especially in the case where there exist over-dispersion in the count data, then the distribution has to be adjusted using a combination of distributions. This is a common application in econometric, ecology and medical biology. Greene (2005) showed how log-normal distri-

bution can be used for log-gamma to model unobserved heterogeneity and Negative binomial used as an introduction of latent heterogeneity to Poisson and/or log-linear count data. Thus if we ignore the possibility of mixture models in multivariate data which normally results in a probabilistic clustering then we should consider specifying a combination of mixed models to cater for the heterogeneity in univariate case.

These normally rise in parametric distribution when certain parameters ceases to be constants due to different factors thus becoming a variable, Boswell et al. (1979), Erol (2014). The mixture model in univariate data generally follows the following distribution:

$$\begin{aligned} f(x) &= \int f(x,p)g(p)dp \ \ continuous\ case \\ f(x) &= \sum_{p} f(x,p)g(p) \ \ discrete\ case \end{aligned} \tag{2.3}$$

Where x represents the random variable following a distribution with pdf of $f(x,p)$, p being the mixing parameter with distribution density function $g(p)$. This normally results in a compound distribution herein called mixed distribution. This model caters for the extra variability. As an example: If a Poisson parameter λ is not considered constant but to follow a distribution we obtain a Poisson mixed model to account for the unobserved heterogeneity.

2.3.2. Heterogeneity and Mixture Models

If we have a population in which more than one distribution is assumed without information which observation goes to which distribution (missing data) homogeneity cannot work, Kessler and McDowell (2012). A mixture of parametric distributions through a FMM can be used to model the data in such cases: estimating the parameters to the distributions and data labeling-estimating probabilities of the membership to each observation, McLachlan and Peel (2000). The FMM also deals with data that is multimodal, skewed, heavy tailed or exhibit kurtosis which cannot be ex-

plained or represented by known standard distributions thus capturing the heterogeneity. Mixture models which are skewed and nonnormal helps in many ways to catch the real features of the data. The skewed t distribution has been used in FMMs to deal with light tailed data in many application whilst heavy tailed data seem to perform better using mixtures of exponential distributions as shown by Dang et al. (2015).

These non normal distributions also help do away with the spurious classes for deviations as demonstrated by Asparouhov and Muthen (2015) when they used structural equations with skewed distributions which then deals with the skewness and kurtosis. Non-parametric methods like kernel density functions can be used. However, they may not be able to give the required unknown distributions. Thus FMM provide an alternative that describes the unknown distribution in terms of the mixtures of known distributions and enables you to simulate draws or assess probabilities from the unknown distribution the same way we do with known distributions.

A model based approach has the advantage of detecting heterogeneity of the latent groups which may differ in their means or factors of within group variations and covariation as given in model function (1.1). Although modeling of differences in variance components in human population via function of trait scores via mixture models has been challenging, FMMs have a given advantage over such challenges, Schlattmann (2009). This is because they offer practical means of inferring into the population heterogeneity when analysing human population data like genetically informative data, Gillespie et al. (2006). Heterogeneity can be observed when parameter estimates between groups is expected prior to phenomenon i.e if it is possible to define subpopulations based on observed variable.

Unobserved heterogeneity is when theory does not assume heterogeneity even though it exists or theory captures it but specified groups do not tend to sufficiently capture it in the population. The variables causing heterogeneity in the data are not known beforehand. Thus significant heterogeneity in the data across unobserved groups often results in biased

parameter estimates and hence invalidity of results, Becker et al. (2013), Lubke and Muthen (2005). Different methods can be used to analyse the data with unobserved heterogeneity. However, the estimation of parameters in the presence of heterogeneity is best made possible via a mixture of distributions thus using FMM. This allows modeling of two or more components which differ in means and covariation to be infered, Gillespie et al. (2006).

The use of FMM identifies the heterogeneous mixture components, behaviour of the subgroups and labels the data which makes it possible to infer to the unknown distribution via the mixtures just like we do with standard known parametric distributions. The latent groups with less dramatic group differences, normally characterised by a flat top density function would be most likely difficult to detect via FMMs.

The mixture may be of the same parametric family or heterogeneous family as in Suhaila et al. (2011), Erisöğlu et al. (2011) respectively. Suhaila et al. (2011) demonstrated how a mixed multinomial with an identical parametric model like gamma, weibull, normal can be use in finite mixture modeling to model rainfall patterns in the different geographical areas considered. Whilst the later showed how heterogeneity can be modeled as a mixture model of two heterogeneous parametric distribution in survival analysis. This was done via a exponential-gamma, exponential-weibull distributions. When we have count data many a times in human populations it is not ideal to assume that the components will assume a Gaussian density functions. This may result in a mixture of mixture models as in Chapter 3 of Schlattmann (2009). Tang et al. (2015) also showed that highly dimensional data can be be clustered using FMM where random block effects are incorporated with effect in each block taken as random variables. In regression unobserved heterogeneity occur in important covariates have been omitted either in the data collection or somehow thus influence not accounted for in analysis. This results in biased parameter estimates. Relaxing the assumption of normality and homogeneity in parameter distribution as shown by Grün and Leish (2007) makes use of mixture models.

The equation (1.1) often represents the FMM in which all the indicators are continuous. Often we are, however, confronted with other types of indicators, like nominal or ordinal variables or counts. FMM cluster models for nominal and ordinal variables assuming (restricted) multinomial distributions for the items are equivalent to standard exploratory FMM as proposed by Goodman (2007), Bockenholt (1993), Karlisa et al. (2007) for Poisson counts. Actually according to Magidson and Vermut (2002), we can specify components for cluster models using the general structure of the FMM model for mixed-model data. Assuming local independence, the FMM cluster model for mixed $y_i s$ is of the form

$$f(y_i|\theta) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} f_k(y_{ij}|\theta_{jk}) \tag{2.4}$$

where J denotes the total number of component and j a particular component. Rather than specifying the joint distribution of y_i given class membership using a single multivariate distribution, we now have to specify the appropriate univariate distribution function for each element y_{ij} of y_i. Possible choices for continuous y_{ij} are univariate normal, student, gamma, and log-normal distributions.

A natural choice for discrete nominal or ordinal variables is the (restricted) multinomial distribution. Suitable distributions for counts are, for instance, Poisson, binomial, or negative binomial, Suhaila et al. (2011), Magidson and Vermut (2002b). FMM also play a vital role in life testing and reliability as shown by Sultan et al. (2007), when they introduced an identifiable mixture model of two inverse Weibull distributions and a mixture of distributions to survival data by Erisöǵlu et al. (2011) which also addressed identification. These FMM with both discrete and continuous parametric distributions can be also used especially in environmental modeling, Suhaila et al. (2011).

In summary FMM can capture heterogeneity in four ways as reviewed by Zhang and Huang (2015):

- Models with normal mixture for independent data,
- Models with normal mixture of longitudinal data,
- Models with non normal independent data like skewed or atypical distributions for real data is non normal,
- Models with non normal longitudinal data

In sum, model based clustering explicitly targets the identification of unobserved heterogeneity in a population based on the observed data, McLachlan and Peel (2000) and utilizes the expectation-maximization (EM) algorithm for maximum-likelihood (ML) estimation. The tasks of clustering become that of estimating: (*i*) the parameters of the assumed mixture (i.e., nature of clusters such as their respective means, covariances) and (*ii*) the posterior probabilities of cluster membership, Raftery and Banfield (1993) by maximizing a log-likelihood function. The convergence (and estimation) is achieved when iterations between E-steps and M-steps result in two successive likelihood values of parameters that differ less than specified in an a priori set threshold. The cluster membership for each case is then determined by assigning each case to the cluster with the highest posterior probability.

2.3.3. Mixture Model Based Clustering

Clustering has been one of the most difficult problems with a large number of heuristic problems in literature. Its a general type of classification in which the classification is not observed and there is little or no prior knowledge of the structure being looked for in the analysis. The problem of searching for group of related observations in a given data set is called cluster analysis, Fraley and Raftery (1998). Although heuristic distance based methods have been used for clustering, one disadvantage of hierarchical clustering algorithms, k-means algorithms an others is that they are largely heuristic and not based on formal models making formal inference impossible.

Model-based cluster analysis is a new clustering procedure to inves-

tigate population heterogeneity utilising finite mixture multivariate normal densities. It is an inferentially based, statistically principled procedure that allows comparison of non-nested models using the Bayesian Information Criterion (BIC) to compare multiple models and identify the optimum number of clusters. Thus model based clustering is based on the idea that instead of coming up with single population, the observed data come from a source with several sub populations. It can be done on either univariate or multivariate data. Early studies by Pearson (1894) on the crabs data was part of the introduction of model based clustering when they used a mixture of two univariate Gaussian distributions to analyse a data set containing ratios of forehead to body lengths for 1,000 crabs, using the method of moments (MOM) to estimate the parameters in the model.

Since then several researches have followed in different disciplines using these mixture model based clustering. Smyth (1999) showed that model based clustering can be done on static characteristic (multivariate data) coupled with dynamic behavior (sequential data) using probabilistic clustering models in a principled and coherent manner. The general idea is to model each sub population differently and overall population as a mixture of these using FMMs. This provides a partition of the data into groups of homogeneous observations. The partitioning is done without prior information or knowledge of the data thus creating a problem which incorporates missing data into the analysis, either by a process specifically designed for missing data or as an integral portion of the standard multivariate analysis, McLachlan and Peel (2000), Hair et al. (1998).

The mixture model is treated as a distribution in which the class labels are missing, but the class labels can be treated as random variables. If these labels were known, we could get closed-form parameters estimates in each component distribution by partitioning the data points into their respective groups. Maximum likelihood estimation techniques is one of the well known process to deal with missing data problem as it make accurate and reasonable estimates. One example which incorporates the likelihood

principle is EM algorithm approach, this is a two stage iterative method with the E and M stages. The E-stage makes best possible estimates of the missing data whilst the M-stage calculates the parameter estimates. Integral approach involves the inclusion of missing data directly into the analysis by defining the missing data as a variable. This makes model based clustering a statistically principled method against other methods due to its ability to determine the best model since it is an inferential procedure.

Model based clustering also uses estimated membership probabilities to classify cases into the appropriate clusters. It uses the assumption that data is generated by a mixture of probability distributions in which each component represents a different cluster. There are two methods used for model based clustering:

- Classification likelihood approach This maximises the

$$L_c(\theta_1,\ \theta_2,.....,\theta_k,\ \gamma_1,\ .....\gamma_k|\mathbf{D}) = \prod_{i=1}^{k} f_{\gamma_i}\left(x_i|\theta_{\gamma_i}\right) \tag{2.5}$$

  where $D = (x_1,\ x_2,\ .....x_n)$ is the set of observations, $f_{\gamma_i}\left(x_i|\theta_{\gamma_i}\right)$ is the density of the observation x_i from component j, γ_j labels the classification with $\gamma_j = j$ if the x_i belongs to the jth component, k is the number of components and θ_j is the corresponding parameters.

- Mixture likelihood approach

$$L_m(\theta_1,\ \theta_2,.....,\theta_k,\ \pi_1,\ .....\pi_k|D) = \prod_{i=1}^{n}\sum_{k}^{j=1} \pi_i f_j(x_i|\theta_j) \tag{2.6}$$

  where π_j are the mixing proportions giving the probability that an observation belongs to the jth component and $\sum_{j=1}^{k}\pi_j = 1$

The most popularly used is the mixture-model clustering, where each latent class represents a hidden cluster, Fraley and Raftery (1998), Magidson and Vermut (2002), McLachlan and Basford (1988), Magidson and Vermut (2003), Vermut (2005). This is versus the traditional clustering approaches which utilises unsupervised learning/classification algorithms that

group cases that are "near" each other according to some ad hoc definition of "distance". The estimation approach in model based clustering is based on probabilities on missing data for the groups using standard methods like EM algorithm, Bayesian approach in parametric FMM.

Although model based clustering first stage make use of the model based agglomerative clustering to determine the initial values of the EM algorithm, it does not use the distances as in the other traditional methods. Advantages of mixture model based clustering includes the following:

- Component distribution have high peaks a sign of tight cluster
- It covers the data so well since all dominant patterns are captured in by component distribution.
- There exist well studied statistical inference techniques to deal with the parameter estimation.
- It is very flexible way since there is freedom in choosing component distributions
- Each component can be estimated by a distribution.
- No linear scale transformations are required
- There is a lot of software classification available.

These are more flexible than the traditional K-means and hierarchical clustering methods. The covariance matrix also give more information on the correlations of the groupings. Mixture Model based approach also allows estimation and hypothesis testing within the framework of standard theory of statistics, McLachlan and Basford (1988). Thus model based clustering provide a principal to deal with several problems like number of components, initial values of parameters and distributions of components.

2.4. Parameter Estimation

2.4.1. Maximum Likelihood Estimators(MLE)

We let **Y** be a p-dimensional random vector with probability density function (p.d.f) $g(y;\ \phi)$ on R^P, where $\phi = (\phi_1,\ \phi_2,\ .....\Phi_d)^T$ is the vector containing the unknown parameters in the postulated form for the p.d.f. of Y. The

superscript T denotes the transpose of a vector or a matrix. The parameter space is denoted by ω. We are going to take $\mathbf{Y}$ to be a continuous random vector, although we can still have $g(y; \Phi)$ as a discrete p.d.f. and adopt counting measures. If the vector ϕ is to be estimated by maximum likelihood. Then likelihood function for ϕ formed from the observed data y is given by

$$L(\phi) = g(\mathbf{y};\ \phi).$$

An estimate ϕ^* of Φ can be obtained as a solution of the likelihood equation

$$\frac{\partial L(\Phi)}{\partial \Phi} = 0 \; or \; \frac{\partial log\, L(\Phi)}{\partial \Phi} = 0 \tag{2.7}$$

This is the general MLE and it can have variations depending on the space being dealt with. Parameters can therefore be estimated by minimising the MLE and solving equation 2.7.

2.4.2. EM Algorithm

Expectation-Maximization algorithm, popularly known as the EM algorithm is a general purpose algorithm for maximum likelihood estimation in a wide variety of situations best described as incomplete-data problems, McLachlan and Peel (2000), McLachlan et al. (1997), McLachlan et al. (1999). The incomplete data situations where the EM algorithm has been successfully applied include not only evidently incomplete data situations, where there are missing data, truncated distributions, censored or grouped observations, but also a whole variety of situations where the incompleteness of the data is not natural or evident. Thus, in some situations, it requires a certain amount of ingenuity on the part of the statistician to formulate the incompleteness in a suitable manner to facilitate the application of the EM algorithm in a computationally profitable manner.

The EM algorithm is a broadly applicable approach to the iterative computation of MLE, useful in a variety of incomplete data problems, where algorithms such as the Newton-Raphson method may turn out to be more

complicated. On each iteration of the EM algorithm, there are two steps-called the expectation step or the E-step and the maximization step or the M-step. Because of this, the algorithm is called the EM algorithm. The basic idea of the EM algorithm is to associate with the given incomplete-data problem, a complete-data problem for which ML estimation is computationally more tractable; for instance, the complete-data problem chosen may yield a closed-form solution to the maximum likelihood estimate (MLE) or may be amenable to MLE computation with a standard computer package. The methodology of the EM algorithm then consists in reformulating the problem in terms of this more easily solved complete-data problem, establishing a relationship between the likelihoods of these two problems, and exploiting the simpler MLE computation.

In the E-step new provisional estimates of the component membership probabilities (i.e., the posterior probabilities) are obtained based upon provisional estimates of Ψ. In the M-step new estimates of Ψ, are obtained on the basis of these provisional estimates of the component membership probabilities. Let $g_c(\mathbf{x};\psi)$ denote the p.d.f. of the random vector χ corresponding to the complete-data vector $\mathbf{x}$. Then the likelihood function that could be formed for ψ if $\mathbf{x}$ were fully observable is given :

$$log\ L_c(\psi) = log\ g_c(\mathbf{x};\psi). \tag{2.8}$$

Formally, we have two samples spaces χ and γ and a many-to-one mapping from χ to γ. Instead of observing the complete-data vector $\mathbf{x}$ in χ, we observe the incomplete-data vector **y = y(x)** in γ. It follows that

$$\mathsf{g}(\mathsf{y};\psi) = \int_{\gamma(\mathsf{y})} \mathsf{g_c}(\mathsf{x};\Psi)\mathsf{dx}, \tag{2.9}$$

where $\gamma(y)$ is the subset of χ determined by the equation $y = y(x)$.

The EM algorithm solves the incomplete-data likelihood of equation (2.1) indirectly in an iterative way using log likelihood function, $log\ L_c(\Psi)$. As it is unobservable, it is replaced by its conditional expectation given y, using

the current fit for ψ. More specifically, let $\psi^{(\mathbf{0})}$ be some initial value for Ψ. Then on the first iteration, the E-step requires the calculation of

$$Q(\Psi, \Psi^{(0)}) = E_{\Psi^{(0)}} \{log\, L_c(\Psi)|y\} \tag{2.10}$$

The M-step requires the maximization of $Q(\Phi, \Psi^{(0)})$ with respect to Ψ over the parameter space ω. That is, we choose $Q(\Psi^{(1)})$ such that $Q(\Psi^{(1)}, \Psi^{(0)}) > Q(\Psi, \Psi^{(0)})$ for all $\Psi \in \omega$. The E- and M-steps are then carried out again, but this time with $\Phi^{(0)}$ replaced by the current fit $\Psi^{(1)}$. On the $(k+1)^{th}$ iteration, the E and M-steps are defined as follows:

E-Step: Calculate $Q(\Psi, \Psi^{(k)})$ where

$$Q(\Psi, \Psi^{(k)}) = E_{\Psi^{(k)}} \{log\, L_c(\Psi)|y\} \tag{2.11}$$

M Step: Choose $\Psi^{(k)}$ to be any value of $\Psi \in \omega$ that maximises $Q(\Psi, \Psi^{(k)})$ i.e

$$Q(\Psi^{(k+1)}, \Psi^{(k)}) > Q(\Psi, \Psi^{(k)})$$

for all $\Psi \in \omega$. The E and M steps are repeated until

$$L(\Psi^{(k+1)}) - L(\Psi^{(k)})$$

changes to an is arbitrary amount for convergence.

2.4.3. Question of Optimality

Optimality actually deals with identifiability issues which seeks to answer if unique solutions can be obtained. There are generally three general concerns: The first concerns whether a distribution or a family of distributions, parameterised in terms of one or more natural parameters, is generally identifiable. The second concerns the estimation procedure used - whether ML estimation yields identifiable solutions. And the third concerns the particular response model being fit and the class of indeterminacies that are relevant, Dillion et al. (1994). The evidence suggests that a large num-

ber of distributions are identifiable. The likelihood function for the EM algorithm is normally non concave and can be maximised by different starting values.

Thus there is need to come up with particular methods to identify the optimal clusters and best ones. This can be determined in so many ways by using different optimal tactics and literature has shown. These involve the following:

1. Parametric Boostrapping. The procedure involves drawing $(t-1)$ random Monte Carlo samples of size N from a population having k-components and pre-specified density, and fitting k and $k+1$ component models to each of the generated samples.
2. Information theoretic based model selection criteria to assess goodness of fit. With this approach the decision on the number of components to retain proceeds by minimizing a particular information-theoretic criterion. This involves the use of the Aikake Information Criteria (AIC), Bayesian information criteria (BIC), Comprehensive Aikake Information Criteria (CAIC) among others.
3. Entropic complexity criterion called ICOMP. Similar to AIC, BIC, and CAIC, ICOMP is a penalized information-based validity functional. It can be viewed in terms of two distinct components. The first component measures the lack of fit of the model, and the second component measures the complexity of the estimated inverse-Fisher information matrix. The preliminary evidence on the performance of the ICOMP measure is quite encouraging. More comprehensive and intense simulations are, however, needed before definitive and generalisable conclusions can be drawn.
4. Minimum information ratio (MIR) introduced by Windham and Cutler (1992). The MIR is the smallest eigenvalue of the Fisher information matrix, for the classified sample. The MIR can be viewed as a measure of the proportion of information about the parameters available without knowledge of the subpopulation membership of the

sample.

5. The nonparametric ML estimation (NPMLE), or what is also is called the flexible support size method since itself is treated as an unknown parameter and P is viewed as completely unknown varying in the set of all probability measures. In essence, with NPMLE a grid of parameter values is specified over which the corresponding population proportions that maximizes the likelihood function are found.
6. Weighted average method by Wei and McNicholas (2015). This chooses a number of close best models and averages them to come up with the possible best model.

The most commonly used is the information criteria method through the use of BIC, AIC, etc. In model-based clustering applications, it is common to fit many mixture models within a family and then report clustering results based only on the best one. Such criterion-based methods of model selection have the general feature that the smaller (or larger, as the case may be) the value of the criterion, the stronger the evidence for the model (i.e., the covariance structure and number of components, as well as number of latent variables if relevant, Fraley and Raftery (1998), Brochado and Martins (2005), McLachlan and Rathnayake (2014). However, research by Wei and McNicholas (2015) showed that we can adopt the many models found from the so called best ones. Rather than throw away all but the best model, we can average multiple models by averaging component membership probabilities or averaging models that are in some sense close to the best one, thereby producing a weighted average of clustering results.

2.5. Density Based Nonparametric Clustering

Density based clustering can be done parametrically as explained by McLachlan and Peel (2000), Fraley (2002) and in particular for heart disease as explained in Mufudza and Erol (2016) as well as nonparametrically by identifying clusters as regions of high density separated by regions of low densities, Hartigan (1975). These parametric mixture models where also

explored and examined in chapter 3 and in particular with regard to Poisson mixture regression models. Whilst nonparametric methods associate the clusters to the regions around the modes of the probability distribution of data, clusters in the model-based parametric approach correspond to the components of a mixture of distributions. Density based clusters normally has an advantage as it allows the cluster shapes to be general although it is usually subject to difficulties in the curse of the dimensionality. Density based clustering also allows high accuracy even in high dimensional data as demonstrated by Rinaldo and Wasserman (2010) via a data based bandwidth selection. The difference merges clearly since the number of the modes in a mixture of distributions does not necessarily match the number of components, Menardi (2011).

Nonparametric clustering methods has been of great interest of late as they bring the flexibility required in cluster determination and inference and for any given data type. There are different approaches to nonparametric clustering which makes use of the nonparametric density estimations and sometimes distances too. Another appealing feature of nonparametric clustering estimation techniques is that, by allowing the data to model the relationships among variables, they are robust to functional form specification and therefore have the ability to detect structure which sometimes remains undetected by traditional parametric estimation techniques, Raccine (2004). Nonparametric approach relies on the idea that clusters are associated with the regions around the modes of the density underlying the data.

Although the well known model based mixture model (2.1) for clustering also explained in the previous sections is well known for easy interpretation and inference much need to be done to improve the robustness and the flexibility of the model (2.1) of the form:

$$f(\mathbf{x}) = \sum_{i}^{n} \pi_j f_j(\mathbf{x}) \tag{2.12}$$

The whole idea of model (2.12) is to assume that each component is identified by f_j a parametric density function, thus shape of each cluster is approximated by the same distribution. The clustering problem is now an approximation of the mixing parameters π_j and parameters associated with the density function f_j. This is done under some conditions which makes the model (2.1) identifiable. The most common used densities has been the Gaussian although a lot of variations and shapes have also been considered including using skewed t distributions to capture for a more flexible way for the shapes of clusters.

However, there is still the need to free individual clusters from a given density shape hence the need to explore clusters free of density assumptions and more flexibility via a nonparametric approach. Another limitation for parametric mixture modeling comes from the sometimes serious disparity between a component and a cluster complying to geometric heuristics, Li et. al (2007). If the cluster shapes do not match the shapes of the density f_j, the parametric mixture approach may face difficulties, a motivation for considering completely assumption free cluster shapes densities via nonparametric formulation. The second challenge may also involve variability in the case of the mixing proportions $\pi_j's$ instead of having them as constants.

Nonparametric clustering involves different approaches although the general idea uses kernel density functions which are a representation of mixture functions. The adaptive bandwidth idea is usually used to model adaptive and changing mixing proportions of the different clusters. Thus a non parametric representation of a model based nonparametric mixture models which replaces model (2.12) can be achieved via clusters with kernel density functions. Researchers which include Li et. al (2007) used these kernel density functions as a mixture densities for clusters to identify a point of local maximum of density called a hilltop using the Modal EM (MEM) to identify clusters. Different clusters were linked via a ridge lines linking two hilltops developed, creating a more accurate way of nonparametric clustering through geometric heuristics. Although clustering can also be done via dis-

similarity distances between data points where data points are the clustered according to the distances. Different measures and methods are then used to determine the distances a limitation of flexibility again in cluster determination.

Flexibility in both cluster determination and diagnostics can be achieved via nonparametric mixture models as they cater for more heterogeneity in a practical way. Practical procedures to cluster diagnosis have been used by different researchers on nonparametric clustering who include Menardi (2011) by use of nonparametric density based silhouettes information to diagnose the quality of clusters. Bonhomme et al. (2016) also used nonparametric estimates for finite mixtures from data on repeated measurements with the aim of advancing statistical inference on such models.

The density based nonparametric clustering methods normally use clusters driven by density estimates which determines associated connected regions. An idea developed by Hartigan [Hartigan (1975)] that: given a density function f on R^d and a positive constant c, a (population) c-cluster is defined as a connected component (i.e., a maximal connected subset) of the level set $f > c$. A notion which can be done via modes of a density function. Clusters can therefore defined as connected components of high regions, Cuevas and Fraiman (1997), Cuevas et. al. (2007), Rinaldo and Wasserman (2010). Unlike the parametric methods nonparametric density based clustering methods do not chose particular density function but normally use kernel density estimate with kernels one one's choice. The Gaussian kernel estimates as so far proved to be the common used estimates as used by Azzalini and Toreni (2007), Azzalini et al. (2014), Menardi et al. (2014), Mallapragada (2010). The general commonly used multidimensional kernel estimator is given before as equation (4.3). This can however, be adjusted to be an adaptive kernel with adaptive bandwidth as shown below in equation (2.13):

$$\hat{f}(x) = \sum_{i=1}^{n} \frac{1}{nh_{i,j}.....h_{i,d}} \prod K\left(\frac{x^j - x_i^j}{h_{i,j}}\right) \tag{2.13}$$

3. FINITE MIXTURE MODELS

3.1. Introduction

In real life, data normally comes as a mixture of variables of continuous and categorical data. This is sometimes referred to as mixed data type. Mixed data also comes with difficulties in handling the data since easiest well defined method deals with either categorical or continuous data. In general, a mixed data can be handled by Poisson and Binomial maximum likelihood fitting in a general linear mixed model (GLM) framework. Mixture models can also handle regression cases when the response for each observation y_{ij} is allowed to depend on x_j of the covariate vector, McLachlan and Peel (2000). These models where the response is count (binomial/Poisson) are also sometimes referred to as count data problems. They appear in many real life settings, such as predicting the number of items in economics, analyzing infectious diseases in health and many modeling topics in the corporate world. Mixture models with categorical data often deals with dispersion problems especially in medical research where often the data is count. Thus there has been increasing interest in modeling count data using the Poisson process, geometric and negative binomial, Greene (2007).

There are generally three assumptions done in dealing with mixed data:

- Categorical variables are assumed independent of each other and continuous feature variables are taken to follow a multivariate normal (each one become univariate normal)
- Categorical variables may arise from thresholds of continuous variables, Everitt and Hand (1981).
- Categorical are treated as conditional Gaussians.

Multivariate mixture models which involve mixed data are normally solved

using finite mixture models in a generalised linear model (GLM) or generalised Linear mixed models (GMM) format. In this work we will concentrate on the finite mixtures of the GLM form. A GLM is made up of a linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} \tag{3.1}$$

and two functions: a link function that describes how the mean, $E(Y_i) = \mu_i$, depends on the linear predictor $g(\mu_i) = \eta_i$ and a variance function that describes how the variance, $var(Yi)$ depends on the mean $var(Y_i) = \Phi V(\mu)$ where the dispersion parameter Φ is a constant. These GLM density functions normally involve the use exponential family distributions such as the Gaussian, Binomial, Gamma Poisson and other distributions with the general density form given by:

$$f(y;\theta,\phi) = \exp\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\} \tag{3.2}$$

where ϕ is the dispersion parameter and θ is the scale parameter. Thus mixture of GLM will be mixture densities of the form of equation (2.1) with $g(\mathbf{x}|\theta_i)$ replaced by distribution of the form (3.2). Thus the heterogeneity is captured both in the link function as well as the components.

The mixture models of GLM form can also be used in regression seem as they can account for more than just dispersion but also variability due to heterogeneity challenges. In statistics and machine learning, mixture modeling which infers random probability measures to assign data samples into clusters (mixture components) has become an important area. This is possible mainly via model based clustering which uses estimated membership probabilities to classify cases into the appropriate clusters. In the case of count data, the numbers of samples assigned to each cluster are counts. However, mixture modeling is not typically considered as a count-modeling problem only but involves continuous as well as mixture variable too. Within the family of mixture models, mixtures of linear regressions have also been

studied extensively, especially when no information about membership of the points assigned to each line was available as alluded by Wang et. al. (2007).

3.2. Finite Mixture Regression Models

In regression unobserved heterogeneity occur if important covariates have been omitted either in the data collection or somehow thus influence not accounted for in analysis. This results in biased parameter estimates. Relaxing the assumption of normality and homogeneity in parameter distribution is quite possible by using mixture models as shown by Grün and Leish (2007) where they showed how a mixture regression model in a GLM context can be formulated . Thus a general model finite mixture regression model is considered to be a mixture of k components where each of the components follows a certain parametric distribution. Every component has a weight and prior probability for an observation to come from the given component and the mixture distribution is given by the weighted sum over the K components. The weights may depend on other variables not provided by the data which are referred to as concomitant and these may not be known or known.

In many instances, age and sex have been used to relate to different segments of a population although they do not perform well in a prior segmented environment, thus we can use the mixture model to model such unobserved heterogeneity, Grün et al. (2008), Dalrymplea et al. (2003). Finite mixture of GLM relaxes the assumption that the regression coefficients and dispersion parameters are the same for all observations. Thus each component is described by a GLM with different linear operator but same error and link function. The general mixture model of a GLM is given by:

$$\begin{aligned} H(y|x,\omega,\ \Psi) &= \sum_{k=1}^{K} \pi_k(\omega,\ \alpha) f_k(y|x,\theta_k) \\ &= \sum_{k=1}^{K} \pi_k(\omega,\ \alpha) \Pi_{d=1}^{D} f_{kd}(y_d|x_d,\theta_{kd}) \end{aligned} \tag{3.3}$$

where $\psi = (\alpha,\ \theta_k,\ k = 1:k)$ is the vector of all parameters for the mixture distribution $H()$, y is the response, x is the predictor, ω is the concomitant variable, f_k is the component specific probability density function and $\theta_k = \theta_{kd},\ (d = 1:D)$ is the component specific parameter. Multivariate variables y are assumed to be dividable into D subsets where the component densities are independent between the subsets, hence component density f_k is given by a product over D densities which are defined for the subset variables y_d and x_d for $d = 1,...,D$. This is also true for mixtures of General Linear Model (GLM) where: the d responses are univariate, $\sum \pi_k(\omega,\ \alpha) = 1$ and $\sum \pi_k(\omega,\ \alpha) > 0$, α parameter of the concomitant model.

The exponential family probability distributions are the commonly used for GLM mixture models and these include Poisson, Negative Binomial and Geometric. Often it is assumed that the component specific densities are from the same parametric family for each component, i.e. $f_k \equiv f$ for notational simplicity, and that the link function is also the same for all components $g_k \equiv g$. In a cluster-wise regression setting this will be an obvious model choice as no priori knowledge about differences in distributional families of the components is available. Another popular extension is to have a so-called concomitant variable model for the prior class probabilities, such that the mixing proportions π_k also depend on a set of explanatory variables (e.g., using a multinomial logit model). In such models substantial changes in the parameter estimates with quite small changes in the mixing model is a possible way to conclude that the random effects doesn't have enough information in the data about the distribution (i.e. there are heterogeneity issues). The component intercept parameter α_k can sometimes be aliased with the model intercept resulting in big standard errors for the intercept. Thus a mixture model seeks to alleviate such problems hence addressing overdispersion and heterogeneity. Therefore each component is described by a generalised linear model with different linear predictor, but the same error distribution and link function.

3.2.1. The Poisson Mixture Regression Model

A Poisson regression model is an example of a GLM in which distribution of the response **Y** with covariate vector **x** is Poisson McLachlan et al. (1997) with density:

$$f(y,\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!} I_A(y) \tag{3.4}$$

In equation (3.4), $E(Y) = \lambda$, the link function $g(\lambda) = log(\lambda) = \beta^T x$, $A = \{0,1,2,.....\}$ is the set of nonnegative integers and $I_A(y)$ is the indicator function, which is one if y belongs to the set A and is zero otherwise. The variance $Var(Y) = E(Y) = \lambda$ and there exist dispersion problems when either $Var(Y) > E(Y)$ (overdispersion) or $Var(Y) < E(Y)$ (underdispersion). In many problems overdispersion is more common than underdispersion. Analysis of count data that is overdispersed relative to the Poisson distribution has been studied extensively, it is also known that application of the standard Poisson model may result in misleading inferences. There is therefore need for alternative methods of analysis to deal with the overdispersion problem. Finite mixture models are appropriate when the extra variability comes from the unobserved heterogeneity of the population, which composes of two or more subgroups mixed in various proportions, McLachlan and Peel (2000), Grün and Leish (2007), Dalrymplea et al. (2003).

In particular, Poisson mixture regression can be fitted to such overdispersed count data in the presence of covariate information. A mixture Poisson regression model assumes the objectives are drawn from a finite mixture Poisson where these distributions differ in intercept and in coefficient of the explanatory variable in the regression component or are suspected to be heterogeneous. In general linear mixture models (GLMM), where the count outcomes are clustered or represent repeated measurements, random effects can be incorporated within the regression model to account for the inherent correlation between observations. Poisson mixture regression models include the Poisson mixture model, Negative Binomial (NB), hurdle, Zero-

inflated Poisson (ZIP) and Zero Inflated Negative Binomial models (ZINB), Zhou and Carin (2013). Particularly the ZIP and hurdle models account for both heterogeneity and the extra zeros in the data.

A finite mixture model based on Poisson distribution may be written as equation (3.4) where

$$f_i(y_i,\theta_{ij}) = \frac{e^{-\lambda_{ij}}(\lambda_{ij})^{y_j}}{y_j!} I_A(y_j) \tag{3.5}$$

In equation (3.5) $\log\lambda_{ij} = \beta_i^T x_j,\ \ i = 1:k$ and for the k component mixture

$$E(Y_j) = \sum_{i=1}^{k} \pi_i \lambda_{ij} \tag{3.6}$$

$$Var(Y_j) = E(var(Y_j|Z_j)) + var(Y_j|Z_j) = E(Y_j) + v_{ij} \tag{3.7}$$

where $\lambda_{ij} = \exp(\beta_i^T x_j)$ is the mean of the jth response condition to its membership of the ith component of the mixture, $v_{ij} = \sum \pi_i \lambda_{ij}^2 - (\sum \pi_i \lambda_{ij})^2$, Z_j is the component indicator vector of zeros and ones with $Z_{ij} = (Z_j)i =$ which one if it is from component i and zero otherwise. $v_{ij} = 0$, iff $\lambda_{1j} = \lambda_{2j} = ...\lambda_{kj}$ i.e if the mean for all the components are the same. Thus mixture models cope better than the homogeneous models. Heterogeneity across individuals in a Poisson mixture regression model is tackled in two ways which includes: formulation in such a way that mean event rate has a discrete mixture distribution (i.e varies across a finite number of unobserved classes) and that the mean event rate within each class as well as depending on the explanatory variables.

The general Poisson mixture regression model then assumes the following:

i) The unobserved mixing process can occupy any of the k components where k is finite and unknown
ii) for each of the observed count y_j, there exist an unobserved random variable Z_j representing component j at which y_j is generated and (Y_j, Z_j) are piecewise independent.
iii) Z_j follows a discrete distribution with k points of support $P(Z_j = j) =$

π_j, with $\sum \pi_j = 1$.

v) Conditional on the Z_j, Y_j follows a Poisson distribution.

Here $y_j,\ j = 1:n$ represent the jth response variable such that $y \approx Po(\lambda_{ij})$ with probability $\pi_i(x)$, which are the mixing weights, where x and z are matrices of covariates and k the number of components in the model. Using logit and log-linear links to model $\pi_i(x)$ and λ_{ij}, which denote the vector of the regression coefficients of the ith component respectively gives:

$$\begin{aligned} logit\left(\pi_i(x)\right) &= log\left(\frac{\pi_i(x)}{1-\pi_i(x)}\right) = x_j^T\beta_j \\ &= (1,\ x_{2k},\ ,\ x_{nk})\left(\beta_{0j},\ (\beta_{1j})^T\right)^T; \\ \log\left(\lambda_{ij}\right) &= z_j^T\alpha_j \end{aligned} \tag{3.8}$$

for $\beta_j,\ \alpha_j$ unknown parameters. Whilst the covariates of the Poisson rates z are restricted to be the same, the covariates in the mixing proportions (x) are not restricted to being the same as those in the Poisson rates (z). In Poisson model overdispersion arises when $V(y_i) > E(y_i)$ and can be modified via a different link function, alternative frequency distribution or both.

- If $k = 1$ we obtain the Poisson regression model reduces to Poisson generalised model if β_{11} is specified to null vector.
- If β_{1i} is specified to zero vector we obtain a k component Poisson mixture model.
- If $k = 2$, and $\lambda_{1k} = 0$, for all j then we obtain a Poisson regression model with extra weight at $y_j = 0$.

3.2.2. EM for GLM Finite Mixture Models

The EM for the Poisson mixture model is based on the complete data likelihood and in this case given by:

$$\log L_c(\psi) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ij}\log\pi_k + \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log f(y_i,\ \theta_{ik},\ \psi_k) \tag{3.9}$$

As explained in the EM algorithm before the E step replaces the unobserved indicator variable with their expected values given in the data $(y_1, , y_n)$ denoting the posterior probabilities of the ith observation belonging to the kth component and is given by:

$$e_{ik} = \frac{\pi_k f(y_i,\ \theta_{ik}, \theta_k)}{\sum_l \pi_l f(y_i,\ \theta_{il}, \theta_l)} \tag{3.10}$$

The M step involves maximisation of the current expected data likelihood for the mixing weights given by:

$$\pi_k^{new} = \frac{\sum_i e_{ik}}{n} = \frac{\frac{\pi_k f(y_i,\ \theta_{ik}, \theta_k)}{\sum_l \pi_l f(y_i,\ \theta_{il}, \theta_l)}}{n} \tag{3.11}$$

The component parameters are then found by the maximum likelihood estimate for the second sum of the expected complete data, thus a solution to the following equation

$$\sum_{i=1}^{n}\sum_{k=1}^{K} e_{ik}\frac{\partial}{\partial\beta}\log f(y_i,\ \theta_{ik},\ \psi_k) = 0 \tag{3.12}$$

3.3. Application: Heart disease prediction

The American Heart Association Statistics (AHAS) 2015 reports that heart diseases also known as cardiovascular diseases are the leading global death accounting for 17.3 million deaths per year a number expected to double by 2030. It is believed they kill 1 in 7 in the US, Mozzaffarian et al (2015). Heart disease includes a number of conditions that influence the heart and not just heart attacks. Heart disease also includes functional problems of the heart such as heart-valve abnormalities, high blood pressure (BP), smoking, diet, cholesterol or irregular heart rhythms. These problems can lead to heart failure, arrhythmias and a host of other problems.

Effective and efficient automated heart disease prediction systems can be beneficial in health-care sector. There has been vast works on heart disease prediction especially using data mining techniques by several re-

searchers in the past years. Although they had different methods of classification and algorithm choice, their focus has been on using different algorithms and combinations of several attributes for intelligent and effective heart attack prediction. In a research by Soni et. al. (2011) they found out that decision trees outperform the Bayesian classification even when sometimes incorporated by clustering which can also be improved by genetic algorithm in both cases.

Associative classification has also been used in an effort to come up with the best system of heart disease prediction in combination with the genetic algorithm, Soni et. al. (2011). This is a result also similar to work by Akhil et al (2012) where J48 a decision tree based algorithm out performs the Naïve Bayes and neutral network on a different data set for heart disease prediction. Our work attempts to present the detailed study on heart disease prediction using mixture model based clustering techniques. Most earlier work on heart disease prediction were based on finding the best algorithm for heart disease prediction using supervised classification methods. However, in this work we concentrate on predicting heart disease using unsupervised classification methods via model based clustering.

3.3.1. Heart Disease and Classification

Numerous studies are available in literature on the establishment of algorithms suitable for heart diagnosis using data mining classification techniques. The focus however, has been more on determining the best algorithm for heart diagnosis using a supervised classification approach where we focus on using finite mixture models for heart disease prediction. The use of FMM enables us to determine the variables that are more important for heart disease for each given group or component. Thus the major risk factors for heart disease can be discovered within and among the given different components. In supervised classification, authors like Bhatla and Jyoti (2012), applied Naives Bayes, neutral networks and decision tress to determine the best algorithm for heart diagnosis. They found out that different techniques

achieved different probabilities for different methods. Among the different classification methods like decision trees, Naïve Bayes and neural networks, it has been shown that decision trees are better classifiers for heart disease diagnosis as explored by Dangare and Apte (2012). Regression trees, ID3 and CART usage showed that heart diseases can be predicted based on the important variables which are chest pain, exercises induced (exang), slope and resting sugar level (restecg) with CART being the best classifier alluded by Fang (2008). In this work we will assume that heterogeneity source is unknown and hence we focus on unsupervised classification and clustering via FMM.

3.3.2. Data

The data used in this work is found in the UCI database for the Cleveland Clinic Foundation for heart disease as given by Detrano. Original data has about 76 raw attributes and 303 observations for each attribute. In this research only 14 attributes were used and they are described as in table 0.1. of section 6.6 in appendix. The data had some missing values which were replaced using the mean response imputation method. In most analysis continuous variables were taken to be continuous variables and sometimes changed to log-normal whilst discrete variables with more than two levels were considered ordinal. The response variable (given as num), is a count of the rate at which an observation is diagnosed with heart disease with zero indication no disease whilst 1,2,3,4 indicates the different levels of the heart disease stage. Our focus, however will be to identify the number of groups in the data when no assumption is made on the distribution of any of the variables using the data. No other information is also assumed or known prior to the cluster investigating methods here to be applied. Thus therefore nonparametric clustering methods become convenient and since we are going to use solely the data to determine to estimate distribution and infer information. Nonparametric mixture models using density estimators such as the kernel will be of paramount relevance.

3.3.3. Dependent Variable Distributions

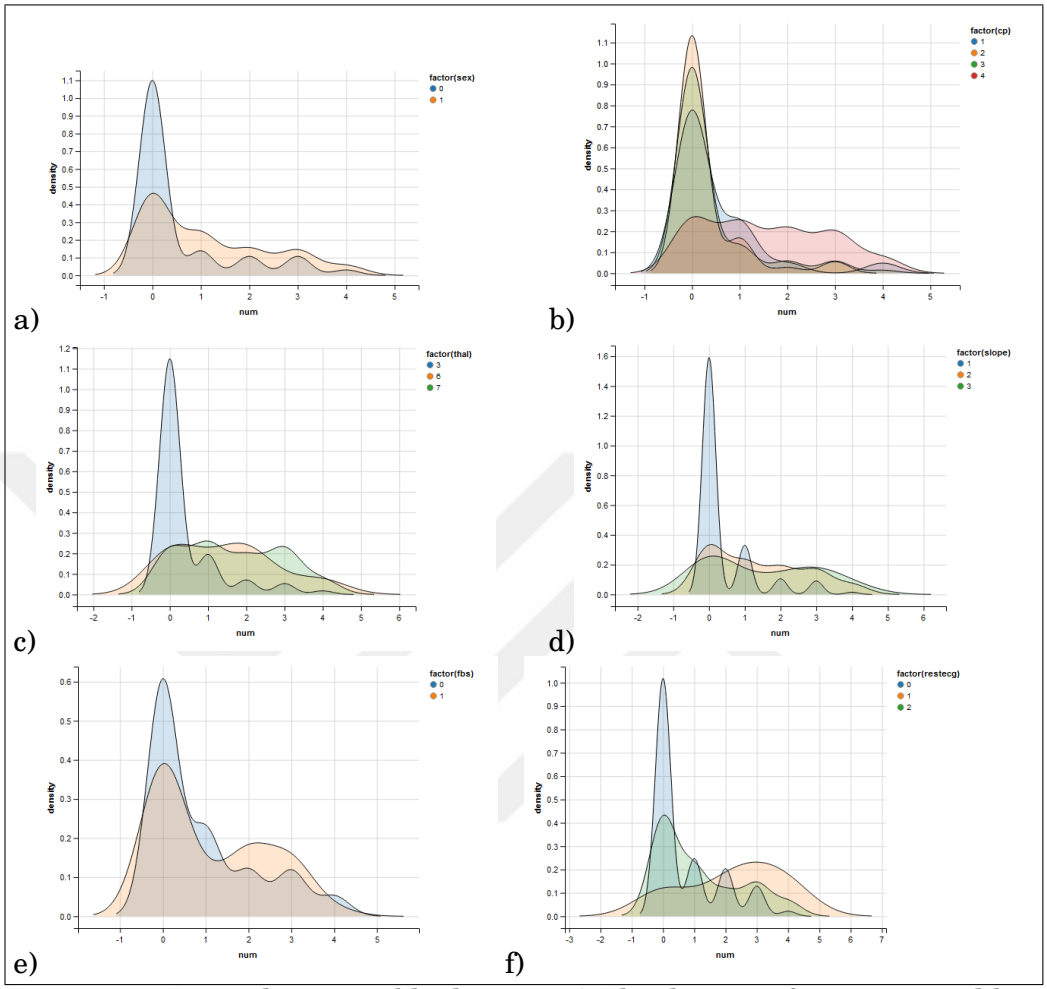


Figure 3.1.: Dependent variable density: a) The density of num grouped by sex, b) density of num grouped by chest pain, c) density of num grouped by thal, d) density of num grouped by slope, e) density of num grouped by fbs, f) density of num grouped by restecg.

Figure (3.1.) shows how the dependent variable changes with sex, cp, thal, slope, fbs and restecg for $(a)-(f)$ respectively. It shows that whilst both males and females densities are positively skewed, more males are affected by heart diseases than females as their rate of diagnosis is higher compared to females. People also with atypical , non angina and asymptotic chest pains have higher changes of being diagnosed with heart disease compared to those with typical angina as shown in figure 3.1.(b). In figure 3.1.(c) shows that individuals with fixed and reversible defects are more prone to heart disease

than those that do not have defects although probabilities of getting such people seems to be slightly less than 0.5. Exercising with up sloping reduces chance of heart disease diagnosis whilst flat and down sloping encourage increase in heart diseases. Although upslopers can still be affected chance are below 0.2 as indicated by figure 3.1.(d). Another heart disease risk trigger is high fasting blood sugar levels and ST-T abnormality as indicated by figure 3.1.(e) and (f), respectively.

Figure 3.2. is a continuation of the dependent variable response

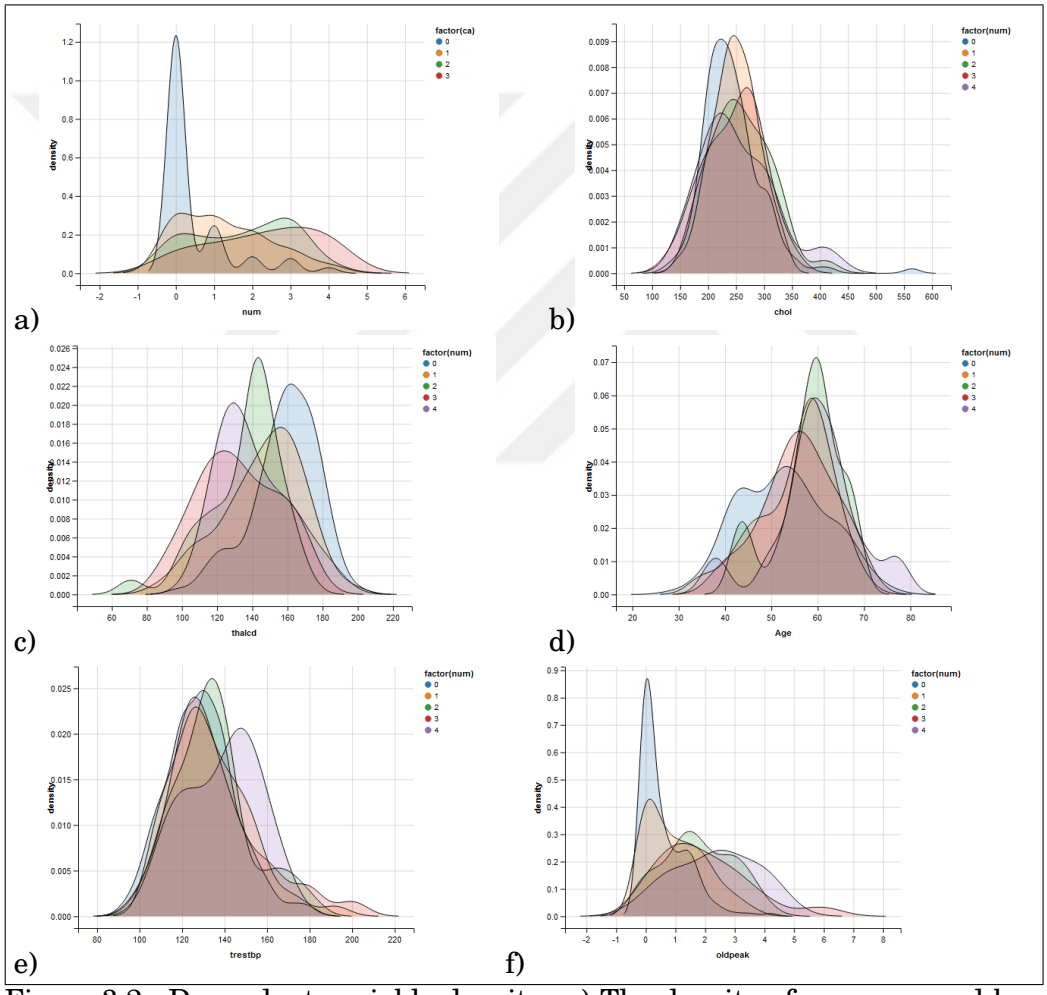


Figure 3.2.: Dependent variable density: a) The density of num grouped by ca, b) density of cholestrol grouped by num, c) density of thalcd grouped by num, d) density of Age grouped by num, e) density of trestbp grouped by num, f) density of oldpeak grouped by num.

with specific variables. It shows that the more the number of major vessels

coloured with flouroscopy the higher the chances of developing heart disease as shown in figure 3.2.(a). Whilst abnormally high levels of cholesterol also triggers high chances of heart disease, the variable seems to follow a fairly normally distributed function as depicted by figure 3.2.(b). A possible high maximum heart rate discourages development of high heart disease rate as in figure 3.2.(c), whilst extremely very high age groups are more prone to heart disease than those in the middle aged by figure 3.2.(d). As expected very high resting blood pressure encourages high heart disease rate like in figure 3.1.(e) and lastly high st induced exercise triggers heart problems as indicated by figure 3.1.(e).

3.4. Generalised Poisson Regression Model (PR)

3.4.1. Why GLM Poisson

Since the data count, a Poisson models will be suitable for the data hence we introduce a GLM Poisson model in this section. As a starting point, the exploration on the data indicated many variations on a Poisson response variable and so a GLM Poisson regression model (i.e. single component model) is introduced as the base model. A GLM is made up of a linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} \tag{3.13}$$

and two functions namely: a link function that describes how the mean, $E(Y_i) = \mu_i$, depends on the linear predictor $g(\mu_i) = \eta_i$ and a variance function that describes how the variance, $var(Y_i)$ depends on the mean $var(Y_i) = \Phi V(\mu)$, where the dispersion parameter Φ is a constant. We are going to use the log function as the link function since our data is Poisson. The exponential family densities (i.e. Normal, Binomial, Gamma and Poisson) distributions commonly used with form densities:

$$f(y;\theta,\phi) = \exp\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\} \tag{3.14}$$

where ϕ is the dispersion parameter and θ is the scale parameter. Mixed data regression problems normally has high complex heterogeneous properties (eg. extradispersion), robustly best handled by mixture regression models in a GLM/GLMM context, McLachlan and Peel (2000).

In this context a GLM with y_i has Poisson response so will have the mean as:

$$\begin{aligned} log(\mu_i) &= \beta_0 + \sum \beta_{j(i)} x_i \\ \mu_i &= \exp\left(\beta_0 + \sum \beta_{j(i)} x_i\right) \end{aligned} \tag{3.15}$$

In this context analysis was done using the heart disease data and a Poisson GLM using the GLM function in **MASS** package in **R**. The results output in table 3.1. show that only sex, chest pain, slope, ca and thal are significant from zero. Thus we have a model:

$$y_i \approx Po\left(\exp\{-3.66 + 0.44sex + 0.32cp + 0.28slope + 0.29ca + 0.18thal\}\right) \tag{3.16}$$

The Poisson regression model indicates that mean heart failure rate

Table 3.1.: Poisson regression Model Output

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.6617599  1.0570249  -3.464 0.000532 ***
Age          0.0068044  0.0084986   0.801 0.423332
sex          0.4162433  0.1744669   2.386 0.017042 *
cp           0.3221903  0.0998690   3.226 0.001255 **
trestbp      0.0026140  0.0036854   0.709 0.478149
chol         0.0002341  0.0011676   0.200 0.841090
fbs         -0.0136699  0.1664693  -0.082 0.934554
restecg      0.1167424  0.0645464   1.809 0.070504 .
thalcd      -0.0052527  0.0032483  -1.617 0.105871
exng         0.2423019  0.1399999   1.731 0.083500 .
oldpeak      0.1020637  0.0563219   1.812 0.069963 .
slope        0.2816738  0.1310286   2.150 0.031578 *
ca           0.2934665  0.0625744   4.690 2.73e-06 ***
thal         0.1785892  0.0410364   4.352 1.35e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Poisson family taken to be 1)

    Null deviance: 511.48  on 302  degrees of freedom
Residual deviance: 249.33  on 289  degrees of freedom
AIC: 628.59
```

depends the multiplicative effects of sex, chest pain, slope of peak exercise segment (slope), number of major vessels (ca) and defects levels (thal). Although the residual deviance proportion to degrees proportion does not seem to be bad, the dispersion parameter is less than one an indication of under-dispersion problems. Thus a GLM Poisson regression model is an inadequate model fit for the data. The Chi- square test on each parameter also shows that most of the parameter estimates are significantly different from zero except cholesterol and fbs as opposed to what the regression model shows. Chi square test on the coefficients only show cholesterol, fbs, and slope are

Table 3.2.: Chi-square test on parameters

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)	
NULL			302	511.48		
Age	1	24.756	301	486.72	6.506e-07	***
sex	1	33.077	300	453.65	8.857e-09	***
cp	1	84.821	299	368.82	< 2.2e-16	***
trestbp	1	11.312	298	357.51	0.0007702	***
chol	1	1.651	297	355.86	0.1988197	
fbs	1	0.521	296	355.34	0.4705207	
restecg	1	4.411	295	350.93	0.0356991	*
thalcd	1	17.508	294	333.42	2.862e-05	***
exng	1	5.413	293	328.01	0.0199927	*
oldpeak	1	29.432	292	298.58	5.791e-08	***
slope	1	2.278	291	296.30	0.1312490	
ca	1	26.599	290	269.70	2.504e-07	***
thal	1	20.371	289	249.33	6.378e-06	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

insignificant from zero with the rest significant however with a lot overdispersion on each of the estimate. The global nullity test also shows that there are at least heart diagnosis is affected by one of the explanatory variables. The Pearson and residual deviances are given in figure 3.3.. It shows both

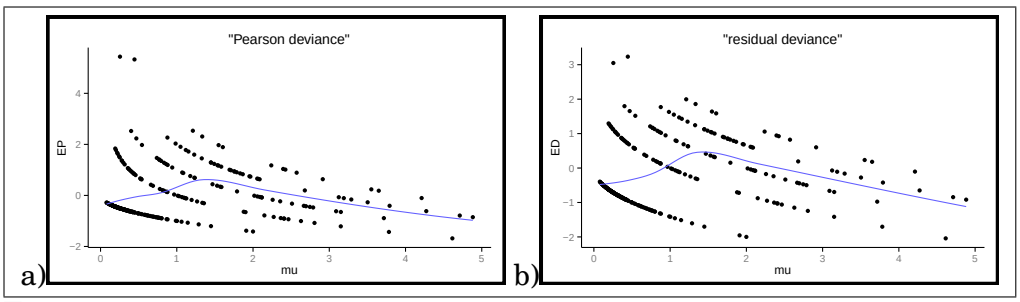


Figure 3.3.: (a) Pearson deviance for the Poisson Regression (b) Residual deviance for the Poisson regression.

the Pearson and residual deviances have well defined pattern on deviances a sign of a poor model fit and also the variance seem to be decreasing as heart diagnosis rate increases.

3.4.2. Residual Plots

Secondly, the regression plots as shown in figure 3.4. also indicates evidently that the Poisson regression model does not fit the data properly as they exhibit heterogeneity. The standard residual plot in figure 3.4.(a) shows that the residual variance is greater than one with a presence of very high leverage points and outliers available in the model. Outliers were on observations 63, 92, 162, 212 and 263. However, 92, 162 and 212 had higher leverages as indicated by 3.4.(d). The QQ plot in figure 3.4.(b) also shows a nonlinear relationship an indication of presence of heterogeneity problems. The residual plot shows that there is deviation in the variance which seems to be decreasing with increase in diagnosis. Hence there is need for a better model which captures the heterogeneity and deviance.

A simpler GLM Poisson regression model with only significant pa-

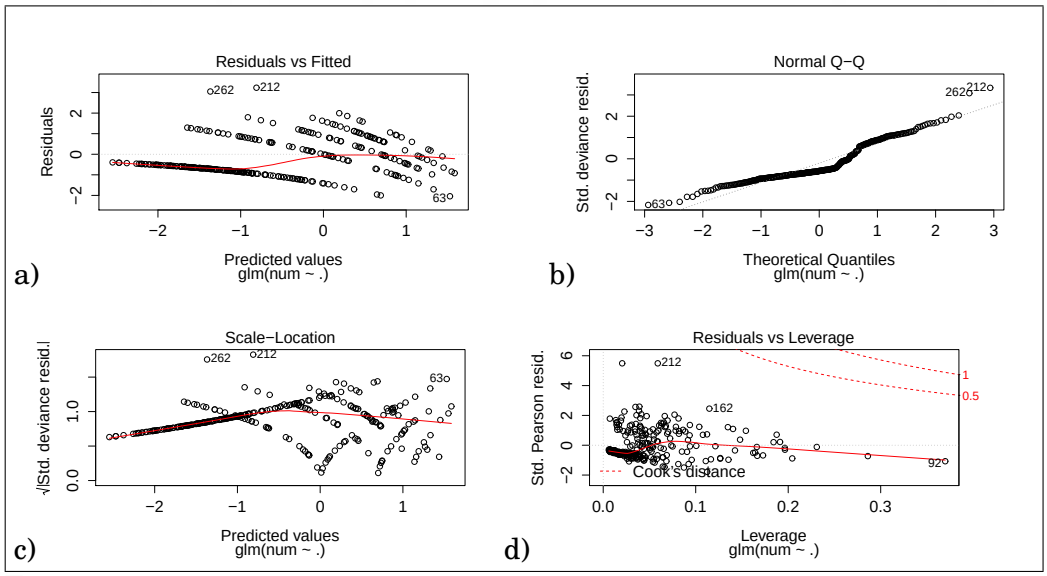


Figure 3.4.: Residual Poisson regression Plots: a) Residual plot, b) QQ plot, c)Deviation and d) Leverage

rameter and general quasi Poisson regression models here not shown were also fitted. Results show that there were no differences between the GLM

Poisson and the quasi Poisson models. Whilst there were a bit of some change in the BIC of the simpler Poisson model due to its simplicity, the rest of the other things still showed out that it was a poor model fit for the data since its AIC was greater than the PR model. The summary of the model fit parameters are shown in table 3.3.. Whilst AIC favour the GLM Poisson model

Table 3.3.: The PR model and Simple PR summary

Model	P.χ^2	Dispersion	D.F	AIC	BIC
GLM Poisson	266.80	0.92	14	628.59	680.58
Quasi Poisson	303.14	1.02	6	630.77	653.05

the BIC favours the simpler Quasi Poisson model. Using the z scaled test for significance shows that the BIC is significant whilst the AIC is insignificant thus there is no difference in the models by the AIC criteria. Although, the simpler quasi model shows that the PR has dispersion close to 1 as in table 3.3., the residual plots and the standard errors for the parameters still shows that it is a poor model fit for the data. Hence a better model other than a Poisson Regression model is required.

3.5. Why Mixture Models on Heart Disease?

Data exploration which has some results shown in appendices indicated that the data has lot of unexplained abnormalities. Most continuous variables showed skewed distributions and a high peak at zero was observed for the response variable with most relationships with the response being non linear. In particular figure 3.1. and 3.2. in section 3.3.3 in the appendix show the existence of groupings within population for most of the variables. Results from GLM Poisson in section 3.4 also reviewed the existence of heterogeneity as it was underdispersed coupled with a lot of unexplained heterogeneity properties as explored shown by the residuals and high leverage outliers. Since, underdispersion is hardly addressed by change of link function or transformations, mixture models tend to be the way to go since they are a holistic approach to heterogeneity. Single parametric distributions will thus not suffice to describe the data in a robust manner no matter how we transform the data.

Therefore a holistic approach which incorporates the irregularities as well as the unobserved heterogeneity within the same model like mixture models does will be required. Mixture models uses a combination of parametric density functions to represent each cluster and the overall population as mixture model thus capturing heterogeneity using several homogeneous groupings. Since a mixture model can capture these by way of dividing the observations into grouped distributions, we seek to employ a finite mixture model to heart disease data with so many explanatory variables causing heart problems and try to capture the heterogeneity in individuals via discrete mixture components. The source heterogeneity is here not known, so we have an incomplete data problem with assumed unknown number of homogeneous groups which will be determined via a finite mixture regression model. Our data has a count response variable and a mixture of both continuous and categorical predictor variables, thus we choose Poisson mixture regression model.

Unlike classification, clustering assumes no prior knowledge is known before hand on the unobserved heterogeneity. Since heterogeneity is in this context going to be tackled via mixture models, we need to be ascertain if the data has clusters using a well defined probability distribution approach for clustering. As such EM- algorithm via Weka will be employed although a different method like K means can also be used. EM algorithm uses probability density functions and prior probabilities in assigning the clusters. This makes it a better clustering method over hierarchical clustering methods like the K means.

3.5.1. EM clustering in Weka

As a quick check to establish if the groupings exist, **WEKA 8.0** a data mining package was used to check for the clusters and have an idea of how many of them using the log likelihood criteria. The method was run for different number of clusters to determine the possible number of components. The results from EM clustering by WEKA are given as table 3.4.: The results

Table 3.4.: EM Results in Weka

K	Cluster distribution	Log likelihood
2	Cluster 1: 56(54%) Cluster 2: 48(46%)	-22.63518
3	Cluster 1: 46(44%) Cluster 2: 25(24%) Cluster 3: 33 (32%)	-25.93119
4	Cluster 1: 20 (19%) Cluster 2: 24 (23%) Cluster 3: 30 (29%) Cluster 4: 30 (29%)	-27.0288

from the EM algorithm by WEKA indicated the best model will have a two clusters model as indicated by the highest log-likelihood which indicates a good fit of the model to the data. In our case since the whole aim is to predict number of diagnosis using the covariates. A regression model which incorporates the clusters and the segmentations within each variable as indicated by data will be more suitable for this heart failure prediction. This calls for a mixture regression model in form of a FMM of a GLM.

3.6. Application: Poisson Mixture Regression (FMPR)

In this section Poisson mixture regression models are applied in order to capture the heterogeneity exhibited by the GLM Poisson models on the heart disease data. Since Mixture models allows grouping of the observations into homogeneous groups we hope to capture for the different variations, deviations and heterogeneity in observations via the Poisson mixture regression model in predicting heart disease. As such a Poisson distribution functions with different parameter values in the different components will be considered as the density distribution function. Thus prediction of the heart disease can be enabled via a finite mixture model with a Poisson function.

3.6.1. FMPR without Covariates

We introduce a GLM mixture model with a Poisson response and log link function without covariates to try cater for the abnormalities exhibited by the normal GLM Poisson model in section 3.4. The mixture model of the GLM Poisson follows the general model in equation 3.4 where f_k is Poisson

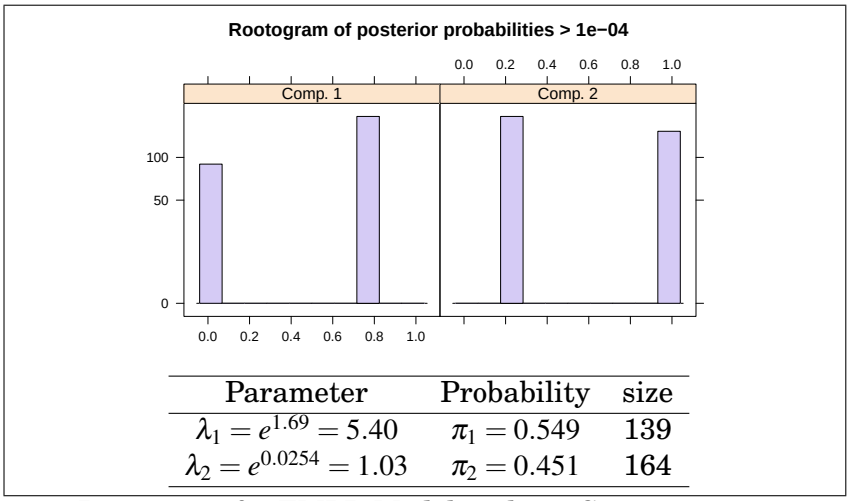
function. This is due to its ability to be able to cluster observations that has correlations into a group as component. The Poisson mixture models the response variable y, given the component membership of the observations are mutually independent. In order to do the analysis, we use **flexmix** package in R by Grün and Leish (2007) and this was done under different number of components given no covariates (i.e $\beta_{1i} = 0$). Table 3.5. shows the results of the Poisson mixture model without covariates under different number of components: Based on both the AIC and BIC a 2 components Poisson mixture model best describes the data when no covariates are considered. We can therefore describe the heart disease diagnosis by a two component Poisson mixture model. Hence we are going to fit a two component Poisson mixture regression model to be able to predict the rate of heart failure given the covariates.

Table 3.5.: Poisson Mixture without Covariates

K	1	2	3	4	5
AIC	864.742	**796.9219**	800.9218	804.9236	808.9137
BIC	868.4558	**808.0631**	819.4905	830.9197	842.3373

3.6.2. Two Component Covariate Poisson Mixture Model

The two component Poisson mixture models was fitted to the data without covariates and some of the model properties and results are given in figure 3.5.. The result shows that those in component 1 seems to have a high heart disease diagnosis rate compared to those in component 2. The average rate at which individuals are diagonised for heart disease in component 1 is higher than those in component 2. The rootogram has peak at probability close to 1 showing that the components are well separated as shown by figure 3.5.. This however, is not a superior model to the Poisson regression model since it has a AIC of 796.92 a higher from 628.59 a suggestion that covariates are very necessary in determining the prediction of the heart disease diagnosis.



Parameter	Probability	size
$\lambda_1 = e^{1.69} = 5.40$	$\pi_1 = 0.549$	139
$\lambda_2 = e^{0.0254} = 1.03$	$\pi_2 = 0.451$	164

Figure 3.5.: Rootogram for FMPR Model without Covariates

3.6.3. FMPR with Covariates

As a mixture model without covariates is not a superior model to GLM Poisson model, we therefore introduce another mixture Poisson regression model with the covariates and analyse how they affect heart disease hence help predict heart disease. The Poisson mixture regression model has the following properties: if we let Y_j denote he j^{th} response, then the heart disease Poisson mixture regression model with 13 covariates assumes that each diagnosis count y_j is associated with at least one of the covariates given. Diagnosis counts are independent and follow a mixture Poisson regression model with conditional means equal to the number of diagnosis. Finally, it also assumes that mean rate for the diagnosis is given by the exponential link function as a multiplicative factor:

$$\lambda_{ij} = exp\{\beta_{0j} + \beta_{1j}x_{1j} + ... + \beta_{kj}x_{kj} \tag{3.17}$$

where x_{ij} represent the covariates. A Poisson mixture regression model was fitted on the heart disease data for both model with extra variation via a multinomial concomitant variable and without the concomitant variable. These models were run under different number of components again using the **flexmix** package in R and the best possible models chosen via BIC crite-

ria. Table 3.6. shows the different models with varying components k, where k_0 is the initial number of components given the covariates and the corresponding AIC , BIC and ICL values for each model. Based on the results from table 3.6., a two component model appears the best models for all the given criteria.

Thus we can say a GLM Poisson mixture regression model with

Table 3.6.: FMPR with Covariate Models

	iter	converged	k	k0	logLik	AIC	BIC	ICL
1	2	TRUE	1	1	-300.30	629	681	681
2	14	TRUE	2	2	-200.67	487	647	656
3	20	TRUE	3	3	-192.77	530	797	854
4	17	TRUE	4	4	-189.31	581	956	1082
5	16	TRUE	5	5	-186.58	633	1116	1310
6	16	TRUE	6	6	-187.22	692	1283	1482
7	21	TRUE	7	7	-186.50	749	1447	1626
8	19	TRUE	8	8	-185.99	806	1612	1953

covariates can be represented as a two component model by both the AIC (487) and BIC (647) criteria for the extra variation model with concomitant ($FMPR_1$) and simple model without the concomitant ($FMPR_2$). This result is also shown graphically as shown by figure 3.6.(a) and (b). Both the model

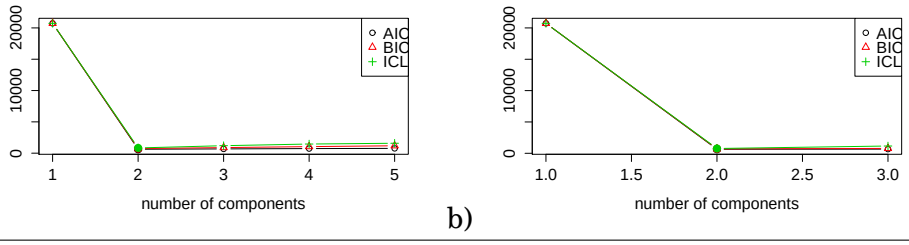


Figure 3.6.: The FMPR Model with Covariates: a) FMPR Models with concomitant variable and b) FMPR Models without concomitant.

with concomitant and without concomitant models heart disease by a 2 component model as evidenced in the figure 3.6..

3.6.4. Properties for Two Component FMPR Model

The two component Poisson mixture models has some of the properties and attributes given as in figure 3.7.. Figure 3.7.(a) and (c) shows the

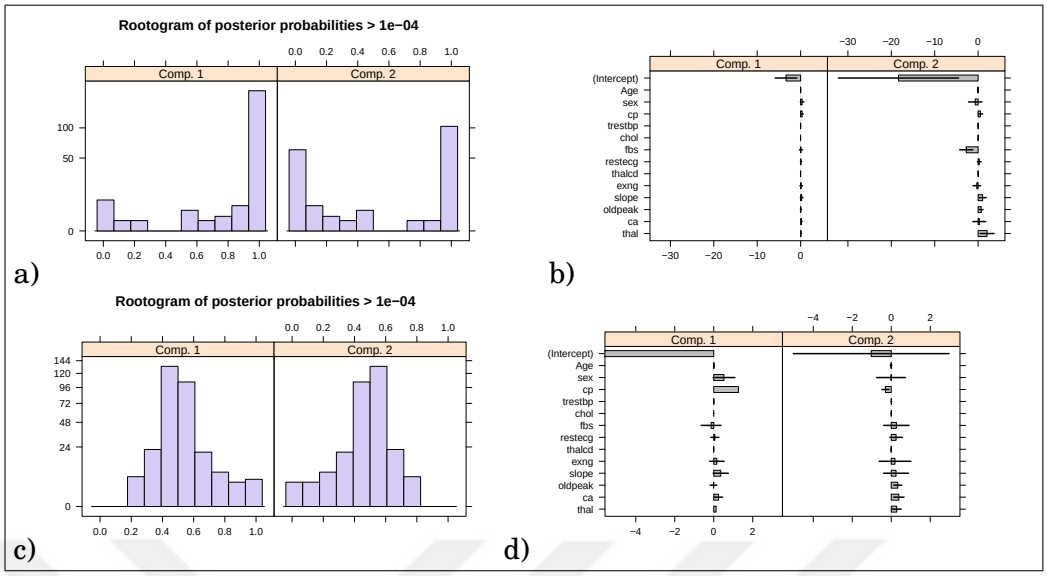


Figure 3.7.: 2 Component FMPR Model with Covariates: a) Rootogram for concomitant variable model b)Parameter Confidence intervals concomitant model, c) Rootogram for model concomitant variable and d) Parameter Confidence Interval for Model without concomitant.

distribution of the components via a rootogram for the model with concomitant variable and without concomitant variable, respectively. The parameter confidence intervals for concomitant model is shown by figure 3.7.(b) whilst (d) represents the non concomitant model. Whilst the concomitant model has well separated clusters due to its highest peak for the distribution being at one, clusters of the non-concomitant model seems to be overlapping as indicated by a fairly separated rootogram which has a lot of mass is within the middle and peak 0.6. There is also a substantial variability within the parameter estimates from the non concomitant model as opposed to the concomitant one as evidenced by wider confidence intervals in figure 3.7.(d) compared to figure 3.7.(b).

In the concomitant model $FMPR_1$ shows that individuals in component 1 are less likely to be affected by heart disease as the parameters likely to be changed whilst component 2 changes with change in any of the covariates given. An indication that individuals in component 2 are at more risk for heart diseases than those in component 1. Major risk for component 2 include sex where males are more affected than females, chest pain where

change from typical angina to asymptomatic may trigger an increased risk to heart disease. Absence of fasting blood sugars reduces chance of heart disease whilst increase in exercise induce angina, number of major vessel, fixed defects and slope exercise increase heart disease possibilities. We can thus conclude that main risk factors include sex, fbs, cp, thal, exang and slope. Non concomitant model ($FMPR_2$) variability is shown in both components and the two components are affected differently by the same risk factors for heart disease.

The results can be summarised as shown in table 3.7.. It shows that

Table 3.7.: Summary: FMPR Models

Model	D.F	AIC	BIC	Prio.Prob.	size
$FMPR_1$	29	654.36	762.06	$\pi_1 = 0.549$	249
				$\pi_2 = 0.451$	54
$FMPR_2$	42	679.31	835.29	$\pi_1 = 0.482$	121
				$\pi_2 = 0518$	182

model with concomitant, $FMPR_1$ is preferred over simplified model, $FMPR_2$ without concomitant by both the BIC and AIC criteria. More observations seem to be classified in component 1 than component 2, where heart disease diagnosis is predicted by increase in most of the variables.

3.6.5. Parameter Estimates

The parameter estimates for the Poisson mixture regression models are the average count rate for diagnosis for each component. These are given as a multiplicative exponential of the significant covariates within each component. The exponents of the parameters estimate values are shown as table 3.8.. The two component FMPR models generally shows that there is a general increase in the heart diagnosis rate as each of the parameters increases or changes. The responses cholesterol (chol), resting blood pressure (trestbp), fbs and maximum heart rate (thalacd) seems to have no effect on the rates of heart disease diagnosis. Considering the $FMPR_1$, a change in one unit increase in heart disease diagnosis seem increase to be 10% more males in component 2 than in component 1 than it does on the females. 30% more

Table 3.8.: Parameter estimates

FMPR with concomitant			FMPR without concomitant		
	Comp.1	Comp.2		Comp.1	Comp.2
coef.(Intercept)	0.02922622	0.02221706	coef.(Intercept)	0.1021142	0.00538968
coef.Age	1.00112515	1.01037506	coef.Age	0.9871701	1.01862162
coef.sex	1.52887074	1.43728449	coef.sex	1.4170347	1.59817604
coef.cp	1.23248085	1.53788848	coef.cp	1.0241905	1.97350239
coef.trestbp	1.00384632	1.00171958	coef.trestbp	1.0042997	1.00259709
coef.chol	1.00005470	1.00036577	coef.chol	1.0018664	0.99895886
coef.fbs	0.98104080	0.97204273	coef.fbs	1.0307847	0.96229211
coef.restecg	1.10443473	1.13360895	coef.restecg	1.1416827	1.12270926
coef.thalcd	0.99436145	0.99509516	coef.thalcd	0.9917247	0.99631471
coef.exng	1.46278919	1.12370738	coef.exng	1.0452589	1.36799965
coef.slope	1.22891730	1.37302521	coef.slope	1.2142637	1.36894104
coef.oldpeak	1.14488269	1.07768620	coef.oldpeak	1.1987679	1.04889725
coef.ca	1.34744345	1.32933631	coef.ca	1.3136992	1.35467881
coef.thal	1.30468218	1.13455558	coef.thal	1.3406171	1.12949769

people are affected by chest pains in component 2 than in component 1 . In general component 1 seems to be less affected than 2 with main risk factors in component 2 being sex, chest pains, exercises, slope and ca more. Thus the heart disease diagnosis for those in different components will be predicted differently. The same can be deduced from the model without concomitant that it clusters differently with cluster one responding to sex, exercise, oldpeak, ca and thal more than component 2 which responds to chest pains , resting blood pressure and slope more than component 1.

However, unlike in the Poisson regression model, the Poisson mixture regression model has most parameters being significant with respect to each of the components. In component 1, ca, thal and cp were significant, whilst component 2 cp, fbs, slope oldpeak and thal were significant. Thus whilst the heart failure for component one is predicted by chest pain, thal and ca , component 2 heart failure prediction is explained by different variables although in both cases chest pain and position defect(thal) determines the diagnosis. Variables like sex, trestbp, thalcd , exang and thal were significant as the concomitant variables an indication that the heterogeneity is being brought by the same variable. Thus component probabilities of the mixture models vary according to sex, resting blood pressure, exercise induced and maximum heart rate of every individual. Therefore, unobserved heterogeneity may have been brought up by unexplained variations within these variables.

3.7. Zero Inflated Poisson Mixture Model

The data used in this work by Detrano involved a lot of zeros in the response variable. We introduce a Poisson model which is suitable for the more zeros in the data through a Zero Inflated Poisson (ZIP) mixture regression models. This is a special mixture model that caters for both heterogeneity and zeros. ZIP model has a mixture of binary distribution that is degenerate at zero and an ordinary count distribution which can be Poisson or Negative Binomial. It considers the zeros to be separate from the non zero, although it can include some zeros in the analysis with non zeros. Zero-inflated models can be thought of as finite mixture models (i.e., where we suppose two data-generating mechanisms, one generating 0âĂŹs and one generating the full range of counts), Hilbe (2014). It has been applied in many problems including epidemiology, occupational health and children growth and development, Fang (2008).

ZIP is used since we have structured zeros within the data which are true zeros from the counting process as well as zeros from dichotomising processes in sampling. The ZIP mixture model assume that

$$y_i = \begin{cases} 0 \; if \; s_i = 0 \\ \\ Poisson \; (\lambda_i, \; \beta, \; \omega) \; if \; s_i = 1 \end{cases} \tag{3.18}$$

Where s_i represents whether component i has heart disease or not. The result from the ZIP mixture model with different components is shown in table 3.9.. The ZIP best model according to AIC, BIC and ICL is the two component model. With the 2 component model having a BIC value of 569.

3.7.1. Two Component ZIPM Model Properties

The 2 component model is here fitted and analysed under 2 different versions: concomitant and nonconcomitant model. Model selection for the ZIPR models was contacted using the following notation: $ZIPR_1$ for model

Table 3.9.: ZIP Models

iter	converged	k	k0	logLik	AIC	BIC	ICL
1	2	TRUE	1 1	-10365.3	20700	20700	20700
2	9	TRUE	2 2	-201.7	461	569	570
3	13	TRUE	3 3	-186.2	488	704	738
4	14	TRUE	4 4	-186.4	547	870	949
5	15	TRUE	5 5	-187.0	605	1040	1140

with concomitant and $ZIPR_2$ for nonconcomitant model. The summary for the models is shown in table 3.10.. It shows that whilst the BIC criteria seems

Table 3.10.: Summary of the ZIPR Models

Model	AIC	BIC	Prior prob.	size
$ZIPR_1$	627.18	731.17	$\pi_1 = 0.3686$	131
			$\pi_2 = 0.632$	172
$ZIPR_2$	645.76	701.47	$\pi_1 = 0.0937$	2
			$\pi_2 = 0.906$	301

to favour a simpler model $ZIPR_2$ model without the concomitant variability, the AIC criteria on the other hand favours a more complicated model $ZIPR_1$. The Zero Inflated Poisson Mixture models also seems to be better models compared to FMPR models due to their lower BIC values.

3.7.2. Cluster Distributions

The component distributions for the 2 component ZIP mixture models can be shown via the rootograms for both models. The component distribution for $ZIPR_1$ and $ZIPR_2$ are shown in figure 3.8.(a) and (b), respectively. The confidence interval for the parameter estimates for non degenerate component for the best model is shown as figure 3.8.(c) and 3.8.(d) shows a graph of the best model selection criteria. The rootogram in 3.8.(a) shows that components for the concomitant model are highly separated. However, in 3.8.(b) components for the simpler model are fairly separated with some overlapping as indicated by the presence of mass in the middle of rootogram. Unlike the finite Poisson mixture regression (FMPR) model, the ZIPR components are better separated as indicated by the their rootograms. Considering the parameter confidence interval for the concomitant model as shown in figure

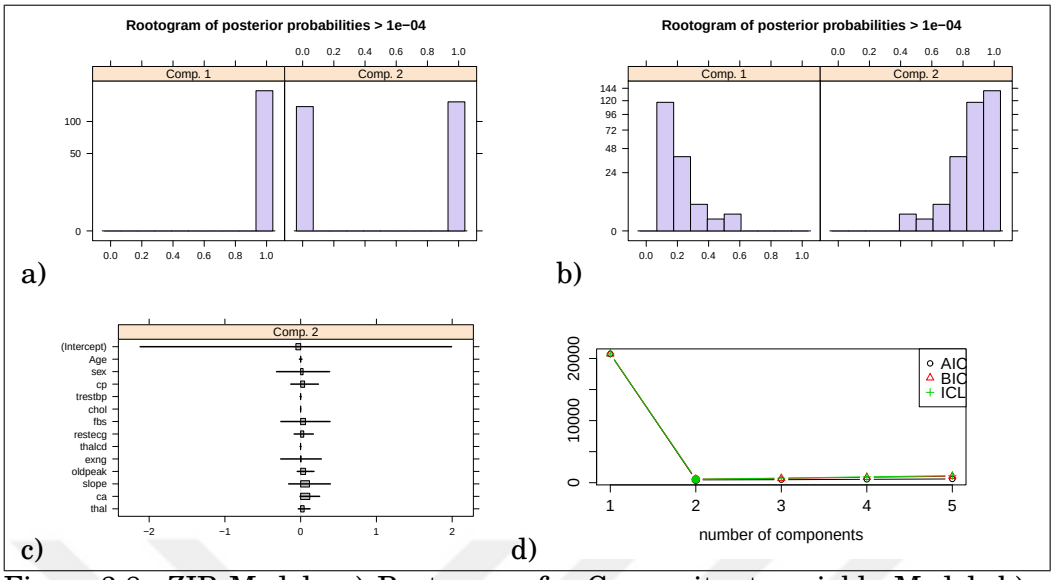


Figure 3.8.: ZIP Model: a) Rootogram for Concomitant variable Model, b) Rootogram Model without concomitant, c) Confidence interval for parameters, and d) Model Selection graphs.

Table 3.11.: Parameter Values of the ZIPR Models

ZIPR with concomitant			ZIPR without concomitant		
	Comp.1	Comp.2		Comp.1	Comp.2
coef.(Intercept)	0	0.120	coef.(Intercept)	0	0.035
coef.Age	1	1.00	coef.Age	1	1.01
coef.sex	1	1.28	coef.sex	1	1.49
coef.cp	1	1.24	coef.cp	1	1.34
coef.trestbp	1	1.00	coef.trestbp	1	1.00
coef.chol	1	1.00	coef.chol	1	1.00
coef.fbs	1	1.04	coef.fbs	1	0.99
coef.restecg	1	1.08	coef.restecg	1	1.10
coef.thalcd	1	0.997	coef.thalcd	1	0.10
coef.exng	1	1.12	coef.exng	1	1.23
coef.slope	1	1.23	coef.slope	1	1.32
coef.oldpeak	1	1.10	coef.oldpeak	1	1.09
coef.ca	1	1.23	coef.ca	1	1.32
coef.thal	1	1.12	coef.thal	1	1.19

3.8.(c), results show that the heart disease can be predicted most variables except resting blood pressure, age, fbs and cholesterol levels.

3.7.3. Parameters

Predictor variables that highly predict the heart disease state can be given as an exponential of model parameters values as shown by table 3.11.. These are given component wise since every component in each different model will predict heart disease in a different manner due to the varying risk factors involved. The ZIPR models show that the first component doesn't

change with any change in the predictors as expected since it is degenerated at zero and hence individual not likely to develop heart disease. However, the second component changes the heart rate diagnosis with changes in the predictor variables. Thus we are able to predict the heart disease diagnosis rate using a mixture of the Poisson regression model with 2 components with a degenerate at zero. The $ZIPR_1$ model predicts the hearts disease through almost all the variables except age, trestbp, cholesterol, and thalcd. The simpler models is not affected by 3 variables including trestbp, cholesterol and fbs.

3.8. Summary

The presence on abnormalities, heterogeneity and extradispersion in count data required a holistic approach to cater for them. GLM Poisson model alone proved useless when there data has underdispersion. However, the use of GLM Mixture model using a Poisson link function proved a better alternative to curb for the underdispersion. A Poisson mixture regression was therefore fitted and analysed both with concomitant and non-concomitant variable. The concomitant proved to be a better model although complex whilst the other is better using the BIC criteria. A zero inflated Poisson Mixture regression model was also fitted to cater for the high frequency zero count of the diagnosis. This degenerate Poisson mixture model proved the the best way to deal with heterogeneity in heart disease as it had the best BIC values and hence best models. Thus we can conclude that heart disease can be predicted using mixture modes were one component represent those who are likely not to develop the disease and the other for those under high risk to develop the disease. In this way mixture models helps to cluster individuals in a heterogeneous population categorically in an effective manner capturing the heterogeneity involved.

4. NONPARAMETRIC MIXTURE MODELS

4.1. Introduction

Data clustering aims to partition data points given in a certain space. There exist no universal rule to define clusters, thus many methods which include both parametric and nonparametric has been proposed. Although parametric methods are very informative and easy to interpret, they include a lot of restrictions and assumptions which can be very misleading resulting in wrong interpretations. These restrictions can be overcome by use of algorithms which make use of nonparametric density estimations. Nonparametric methods limit the number of assumptions made and derive their strength from the sample data given and thus reducing the parametric deficiency. These can be represented by histograms or density estimation methods which include the kernel density estimators. Thus nonparametric clusters are derived from own cluster density functions, unknown but estimated rather than assumed from shape and distribution of a cluster. It is in view of this aspect of improved flexibility that we are going to focus on nonparametric clustering methods.

4.2. Nonparametric Density Estimations

Nonparametric density estimations can be done on both univariate and multivariate data. However, we are going to concentrate on multivariate cases since we will be using multivariate data sets. There are many methods used for density estimations, which include polynomials, kernels, smoothers and many more depending on choice of researcher. However, our concentration will be on kernel density estimators (KDE) as they are applicable to any type of data making multivariate KDE the most commonly used nonparametric density estimators.

4.2.1. Kernel Density Estimation (KDE)

The KDE is compiled such that suppose we have i.i.d d-dimensional sample vectors $\mathbf{X} = (\mathbf{X_1},;\mathbf{X_2},;....,\mathbf{X_n})$ from a population with some unknown probability density f. Then for $\mathbf{X_i} = (X_{i1},X_{i2},;.....,;X_{id}),;i = 1:n$, the multivariate KDE is given by:

$$\hat{f}(\mathbf{x}) = \frac{1}{n|\mathbf{H}|}\sum K(\mathbf{H^{-1}}(\mathbf{x}-\mathbf{X_i})) \tag{4.1}$$

where $K(.)$, is a real-valued multivariate kernel function, such that

1. $\int K(z)dz = 1$,
2. $\int zK(z)dz = 0$ and
3. $\int zz^z K(z)dz = \mu_2(K)\mathbf{I_p}$.

The normally chosen $K(.)$ includes standard multivariate normal density, triangular and Dirichlet functions. H is the $d*d$ non-singular positive definite bandwidth matrix and $|\mathbf{H}|$ is the determinant of H. The d-dimensional H contains $d(d+1)/2$ number of parameters which makes it difficult to choose the values of H due to such a large number of parameters over d-dimensional space. It is more practical to keep the number of bandwidth parameters as small as possible, but retaining enough to provide good estimates. One approach to reduce the number of bandwidth parameters is using simplest model which contains only one bandwidth parameter:

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^d}\sum K\left(\frac{(\mathbf{x}-\mathbf{X_i})}{h}\right) \tag{4.2}$$

However, if the data has different scales on different dimensions, KDE form of equation (4.2) will lead to poor estimations. Several authors in literature who include Cheng and Ray (2014), Härdle (1990) have tried to explain best bandwidth selection methods.

A common formulation for the multidimensional kernel estimator by Scott and Sain (2005) is given by the product kernel estimator:

$$\hat{f}(x) = \sum_{i=1}^{n} \frac{1}{nh_1.....h_d} \prod_{j=1}^{k} K\left(\frac{x^{(j)} - x_i^{(j)}}{h_j}\right) \tag{4.3}$$

where x^j is the jth component of x, the univariate kernel K is usually taken to be a nonnegative function centered at zero and integrating to one, and a different smoothing parameter h_j is chosen for each component.

4.2.2. Kernels

Definition 4.2.1 (Kernel)**.** *It is a mere smoothing or weight assigning real function that is bounded, symmetric and integrates to 1. Thus is a general smooth function such that* $K(x) > 0$ *and*

$$\int K(x) = 1, \ \int xK(x) = 0, \ \sigma_k^2 = \int x^2 K(x)dx > 0$$

The general kernel probability density function (kdf) estimator is given as:

$$pdf(x_0) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right) \tag{4.4}$$

In kernel density estimation, unlike in the histogram, bins are allowed to overlap and place a diminishing weight on the data points as they move further away from point of estimation x_0. The shape of the kernel weights is determined by K, whereas the size of the weights is parameterised by h, the bandwidth. There are so many kernels which can be used but the commonly used estimator is refereed to as the Nadaraya-Watson (NW) kernel estimator. This has many varieties which include Gaussian, Quadratic, Triangular, Epanechinikov, given by the following functions. The kernels in table 4.1.

Table 4.1.: Kernel weights

Type	**Kernel Function**
Rectangular	$K(x) = \frac{1}{2} I(\|x\| < 1)$
Gaussian	$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$
Epanechnikov	$K(x) = \frac{3}{4}(1 - x^2) I(\|x\| < 1)$
Triangular	$K(x) = \frac{70}{81}(1 - \|x\|^3)^3 I(\|x\| < 1)$

does not necessarily change the shape of the density function irregardless of their differences. Although the commonly used is the Gaussian kernel, any other kernel function of any choice can be used.

4.2.3. Kernel Density Estimations Selection

In this section we used Cleveland Clinic foundation data on heart disease provided by Detrano to show how different kernels and bandwidth affect the density functions approximations. The original data consists of 74 variables and 303 observations. However in this work only 14 variables were considered as described in table 0.1. of appendix. Figure 4.1., shows the density distributions of the number of diagnosis, here considered as the dependent variable of the Cleveland heart disease data under different kernels and bandwidths. It shows from figure 4.1.(a) that the shape doesn't necessarily change with change in the kernel function $K(.)$ whilst there is a very big change when a different bandwidth parameter are used as indicated in figure 4.1.(b). The density becomes smoother with increase in bandwidth (h), a process known as over fitting since there is limited bias with high variance. Under fitting on the other side occurs with decrease in h, with increase in the bias and very small variance. As such a trade off between the bias and variance is to be well taken care of and this requires a very educated choice of the bandwidth when using kernel density estimators.

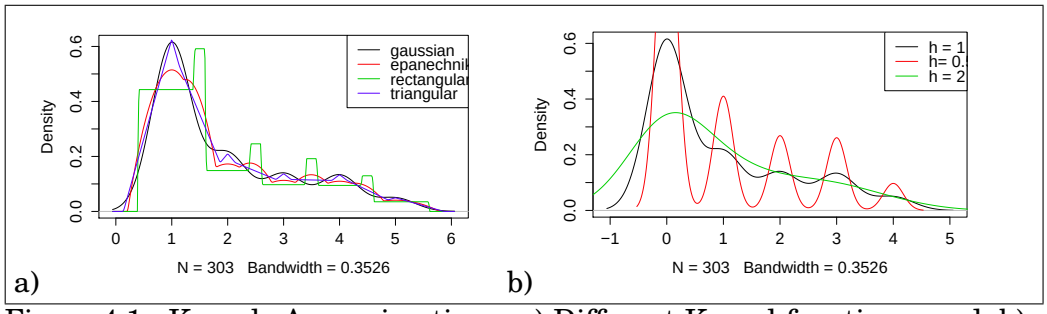


Figure 4.1.: Kernels Approximations: a) Different Kernel functions used, b) Different bandwidth for Gaussian kernel

4.3. Clustering via Nonparametric Mixtures

The general use of nonparametric density based clustering represent a nonparametric mixture model. However, several authors have represented nonparametric mixture models in different ways using general assumption that each cluster is generated by its own unknown density function, Mallapragada (2010). Nonparametric mixture models can therefore also means that no assumptions are made about the form of the $fj's$, even though the weights π_j maybe scalars parameters from model (2.1) as alluded by Chauveau (2015). It should however, be noted that the weights are not only restricted to scalar weights but can be variables. Lindsay et al. (1995), defines nonparametric mixture modeling in a different sense such that the family F from which the component densities come is fully specified up to a parameter θ, but the mixing distribution from which they are drawn is assumed to be completely unspecified and unknown, rather than having finite support of known cardinality.

Nonparametric mixtures models can therefore be used to describe the case in which no assumptions are made about the distribution form the mixture model, even though the parameter mixing parameter can be Euclidean. Semi parametric models normally refers to cases where the distribution form is partly specified by a finite-valued parameter, such as the case in which $f_j(x) = f(x-\mu_j)$ for a symmetric but otherwise completely unspecified density $f()$, as proposed by Hall and Zhou (2003) an idea used by Benaglia et al. (2009b), Chauveau (2015) for the multivariate cases. The assumption that component distributions come from a family of densities that may be indexed by a finite-dimensional parameter vector is normally ignored when dealing with nonparametric mixtures. It is still however, necessary to restrict the family of multivariate density functions from which the component densities are drawn in order to avoid the problem of model non-identifiability, as discussed by Benaglia et al. (2009a).

A number of nonparametric mixture models have considered that the observed variables are jointly conditionally independent given the latent

class and use kernel methods to identify the finite mixtures of nonparametric distributions, Sgouritsa et al. (2013). Suppose θ is a vector of parameters, including the mixing proportions $\lambda_1, ..., \lambda_k$ and the univariate densities f_{jk} where, j indexes the component and k indexes the coordinate, so $1 < k < r$ and $1 < j < m$. Thus, under the assumption of conditional independence, the nonparametric mixture density evaluated at $x_j = (x_{i1}, ..., x_{ir})^t$ is given as equation 1.5. that is

$$f_\theta(\mathbf{x_i}) = \sum_{j=1}^{m} \lambda_j \left\{ \prod_{k=1}^{r} f_{jk}(x_{ik}) \right\} \tag{4.5}$$

Where parameters are explained as in equation (1.5). It should be noted that $\mathbf{x_i}$ sometimes are not only conditionally independent but identically distributed as well which can be represented in terms of blocks as eluded in the work of Benaglia et al. (2009a), where each block represents coordinates that follows the same univariate distribution.

4.3.1. Nonparametric Mixture Model Identification

The curse of dimensionality in mixture models both parametric and nonparametric especially in the multivariate case (1.5) has been a dominant problem. Most researches show that the difficulties comes with dimensions greater than 3, hence conditions needs to be set down such that the parameters are identifiable. Examples include the case of univariate case as alluded by Hunter et. al. (2007) who found that when $f_j(x) = f(x - \mu_j)$ for some density $f()$ that is symmetric about zero, the mixture density can admits a unique representation whenever $m \leq 3$, except in easily enumerable special cases. Curse of dimensionality refers to the question of uniquely identifying the parameter space θ in the following:

- the number of mixture components, m
- the distribution of each component, f_{kj} and
- the mixing weights, λ_k

Addressing this problem involves making different assumption to the mix-

ture model 1.5. However, the question to answer is then how restrictive should the assumption be for uniqueness. In the case of multivariate data, Hall and Zhou (2003) proved that when $k \leq 3$ the problem is parametrically identifiable under particularly mild regularity conditions based on the principle of EM algorithm. A root-n consistent nonparametric mixture model was also identified under joint conditional independence assumption on a Hilbert space condition made by Sgouritsa et al. (2013), some assume parametric family distributions for the components and a discretisation of the observed variables which provides only a lower bound on the number of components was also used by Kasahara et al. (2010).

4.3.2. EM for Multivariate Nonparametric Mixtures

The use of EM like algorithm to estimate parameters of the model (4.5) was introduced by Benaglia et al. (2009a) as a generalisation of the nonparametric EM algorithm via **npEM** algorithm in R. Assuming conditionally independence of the coordinate $\mathbf{X_i}$ vector (independent on conditional of the component/subpopulation from which it is withdrawn). If we allow coordinates to be independent also on existence of blocks that are identically distributed the equation (4.5) can be written as

$$g_\theta(\mathbf{x}_i) = \sum_{j=1}^{m} \lambda_j \prod_{k=1}^{r} f_{jb_k}(x_{ik}) \tag{4.6}$$

Where b_k denotes the block to which the kth cordinate belongs $1 < b_k < B$. The EM algorithm has the E step and M step described in section (2.2) of Benaglia et al. (2009a) where the E step calculates the posterior probabilities of observations conditional on the data of the component inclusion given by:

$$p_{ij}^t = P_\theta(Z_{ij} = 1|\mathbf{x_i}) = \frac{\lambda_j^t \phi_j^t(\mathbf{x_i})}{\sum_{s=1}^{m} \lambda_s^t \phi_s^t(\mathbf{x_i})} \tag{4.7}$$

for $i = 1 : n$ and $j = 1 : m$, where equation (4.7) can be applied as a variety. The M step set is obtained by :

$$\lambda_{ij}^{t+1} = \frac{1}{n} \sum_{i=1}^{n} p_{ij}^{t}, for j = 1 : m \tag{4.8}$$

Where $\phi_j^t(\mathbf{x_i}) = \prod_k f_{jb_k}^t(x_{ik})$ and f_{jl} is obtained by the weighted kernel density estimate given by equation 4.9. Given any component $j \in 1 : m$ and block $l \in 1 : B$

$$f_{jl}^{t+1}(U) = \frac{1}{nh_{jl}C_l\lambda_j^{t+1}} \sum_{k=1}^{r} \sum_{i=1}^{n} p_{ij}^{t} I\{b_k = l\} K\left(\frac{u = x_{ik}}{h_{jl}}\right) \tag{4.9}$$

where $K(.)$ is the kernel density function, h_{jl} is the bandwidth of the jth component and l^{th} block density estimate, and C_l i the number of coordinates of the lth block.

Different authors including Mallapragada (2010) have considered nonparametric mixture models different with different parameter estimation methods. Although they still used the assumption of conditional density function and used KDE for density approximations they constructed profile matrix for each stage to help provide with cluster evaluations. The flexibility of nonparametric density based clustering models has been either on the mixing proportions only, or the mixing proportions as wells as the regression functions which some has employed the use of nonparametric methods for estimation. Although successful the EM like algorithm by Benaglia et al. (2009a) lacks mathematical justification and hence the improvements by Chauveau (2016) by allowing model (1.5) conditionally independent multivariate and nonparametric component densities. Importantly, this extension seems to have allowed dependence structures within multivariate subsets of coordinates, apart from the dependence induced by the mixture that is the unique dependence allowed in model (1.5). They proposed a new EM-like algorithm called mvnpEM which relies on and is a multivariate (mv) per block extension of the npEM algorithm introduced by Benaglia et al. (2009b). The

model is given as (4.10):

$$g_\theta(\mathbf{x_i}) = \sum_{j=1}^{m} \lambda_j \left\{ \prod_{l=1}^{B} f_{jl}(x_{isl}) \right\} \tag{4.10}$$

where $x_{isl} = x_{ik}, k \in sl$ is the multivariate variable which has its coordinates in lth block, thus this model assumes independence of blocks of multivariate densities, conditional on the subpopulation from which each observation is drawn.

This is a main difference in comparison with the model of conditional independence (1.5) introduced by Hall and Zhou (2003). The model bring more flexibility since the dependence structure does not come only from the mixture structure, but from the block as additional within-block dependence is allowed. When all blocks are of size 1 (univariate blocks), then $B = r$ and the model is the conditional independence assumption model (1.5). Having a single multivariate block of size < 3, we assume $B < r$ in the sequel. Block in this sense have a different meaning in Benaglia et al. (2009b) where it previously means a group of coordinates sharing a same univariate density f_j for component j. The EM like algorithm runs pretty like the previous one except the inclusion of the modification as highlighted in Chauveau (2016).

4.4. Nonparametric Mixtures via Graph Theory

Nonparametric mixtures can also be done using graph theory methods where clusters are connected components and estimated using graph based algorithms. The number of clusters can then be chosen based on nearest neighbourhood graph $X_i : \hat{K_h} \geq k$, where K_h is the kernel density estimator and k is the number of clusters. The number of clusters is normally done on the basis of the number of connected components of a level set $f > c$ as explained by Cuevas et al. (2000). Cadre et al. (2007) proposed a combined KDE with single linkage graph to estimate the number of clusters. A lot more authors who include, Rinaldo and Wasserman (2010) have also used density clustering via connected regions using data driven bandwidth selec-

tion measures and stability of the identified clusters. In this sense mixtures are formed by KDE's where an edge between two points indicate that they belong to the same ball ρ, in the same cluster. The number of connected components will then corresponds to the number of clusters found in the sample. Rinaldo et al. (2012) used the same idea when they investigated the stability of the density based clustering by altering different tuning parameters to the mixture models including the kernels.

Graph theory can also be used to identify cluster tress and pruning the trees to do away with spurious clusters in cases where they exist. Stuetzle and Nugent (2010) explored a plug-in approach to cluster tree estimation: estimate the cluster tree of the feature density by the cluster tree of a density estimate . In Azzalini and Toreni (2007) they also describes nonparametric mixture models differently in an interesting way via graph theory by use of the KDE. They identified clusters using connected points by observing the density function along a given interval in a multidimensional space. Clusters are then identified by the existence of some disconnections between Voronoi diagram indicated through a valley using a mode function which is associated with both components and probability. Thus points of high connections indicates a mode of the kernel density and hence a single cluster. Connected regions are then identified using Delaunay diagrams or pair in a high dimensional space, Azzalini et al. (2014). They demonstrated this via the **PdfCluster** package in R as described in Azzalini et al. (2014), Menardi et al. (2014), an algorithm which makes use of the idea by Menardi (2011) where density based silhouette information is used to differentiate clusters and inferences. Thus nonparametric mixture models has that flexibility to identify clusters effectively via so many methods and algorithms, a reason behind our focus on exploring these methods on real data set to help alleviate cardiograhical problems.

4.4.1. Graph Theory Procedure

The method follows a description by Azzalini and Toreni (2007) that supposed if the estimated function is unimodal then the we have a connected cluster otherwise it is disconnected. Thus, if we have $\chi = (x_1, ..., x_n)$, $x_i \in \Re^d$ to be clustered in a d dimensional space with unknown bounded and differentiable density function, f. Then for each constant c, $R(c) = \{x : x \in \Re^d, f(x) \geq c\}$. Azzalini and Toreni (2007) introduced a mode function described such that the function f is replaced by its nonparametric estimate $\hat{f}$. Therefore in multidimensional space where its not obvious to determine the connected region, the interest is to focus on the sample set of $\aleph$ such that we have

$$S(c) = \{x : x \in \Re^d, \hat{f}(x) \geq c\} \tag{4.11}$$

Thus, a multidimensional problem of identifying connected regions is formed. Graph theory is then used to identify connected components as detection of connected components of a given graph **G** whose elements are vertices of $S(c)$ where the edge is key. The task is then done via Voronoi diagram using either Delaunay triangulation for $d \leq 3$ and pairing for $d > 3$ and according to some measure of distance. The Voronoi diagram is a partition of $\Re^d$ given χ into k regions $\Upsilon(x_1), ..., \Upsilon(x_n)$ where $\Upsilon(x_i)$ is the set of all points of $\Re^d$ closer to x_i than to any other point in the set according to some measure of distance. The connected components are identified as union of connected pairs that share at least one vertex and in this way cluster cores are determined. The clusters cores are formed by data lying in the regions around the detected modes. Clusters can also be detected by a minimum spanning tree as explained by Gower and Ross (1969) which is a subgraph of the Delaunay triangulation as explained by Menardi et al. (2014).

Example of how the cluster cores are formed is shown by figure 4.2.. The density estimates which determines the connected regions are not linked to particular density function but are purely estimated. The commonly used density estimate is the multidimensional kernel density estimate by Scott

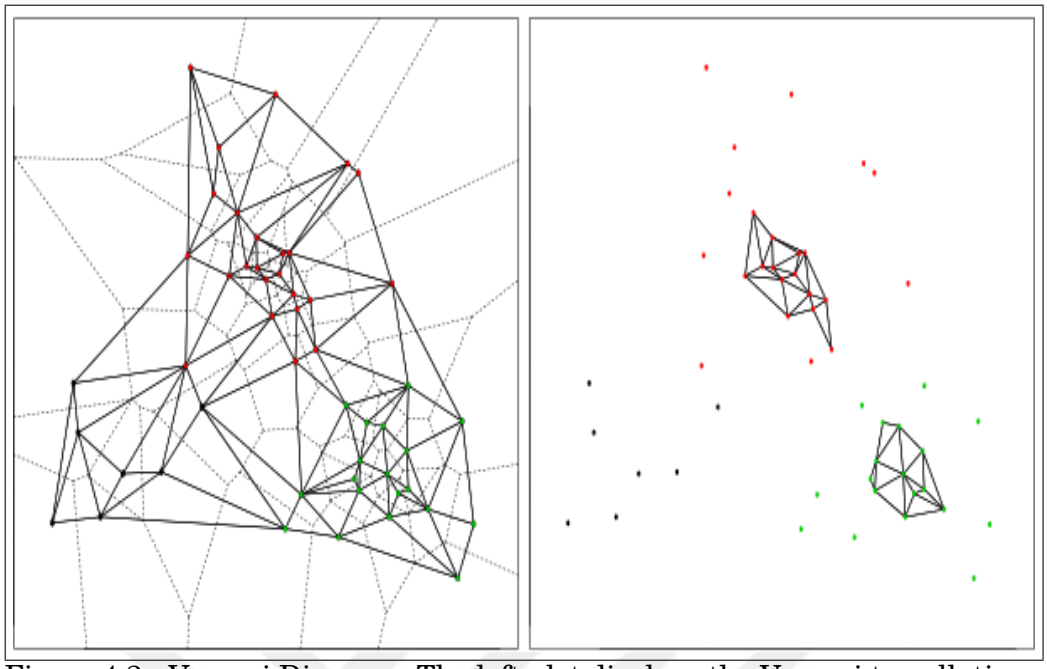


Figure 4.2.: Voronoi Diagram: The left plot displays the Voronoi tessellation (dashed lines) for the points and superimposed Delaunay triangulation (solid lines). In the right plot some edges have been removed from the Delaunay triangulation, those connecting pairs of observations with density below the threshold c represented by the plane in the right plot

and Sain (2005) as given in equation (4.3). We noticed earlier that the choice of the kernel normally does not have an effect on the estimate unlike the smoothing parameter, can be fixed and adaptive as suggested by Stuetzle and Nugent (2010). In this work both the fixed and adaptive bandwidth were used.

4.4.2. Semiparametric Mixture Modeling

Semiparametric Density based clustering (mixture modeling) describes the case where probabilities densities are assumed to be coming from a certain distribution. However, the mixing proportions are assumed to be unknown and variables which also follows some distribution. This random effect in the semi parametric model also accounts for a lot of heterogeneity including the within-cluster correlation. Model-based clustering in the case of dissimilarity data is the statistical approach of fitting the closest classifi-

cation matrix to data. A classification matrix is a clustering model for dissimilarities, i.e., a special dissimilarity matrix with a semiparametric form representing a partition described by two characteristics: heterogeneity for each class and isolation between classes, Vichi (2008). Semiparametric cluster models can also be introduced to account for varying impacts of factors over clusters by using cluster-level covariates as indicated by Zhang et al. (2009). Over the years there has been improvement on nonparametric methods due to the improvement in technology too.

Some of semiparametric models has been studied in different disciplines as improvement from parametric model based clustering methods includes a semiparametric model-based clustering analysis for DNA microarray data. A clustering approach based on mode identification by applying new optimization techniques to a nonparametric density estimator was developed by Li et. al (2007). Mixtures of regressions using semiparametric approach have also been incorporated by authors like Hunter et al. (2012) and Huang et al. (2013). Bayesian semiparametric logit models have also been fitted to grouped data related to in-hospital survival outcome of patients hospitalized with an ST -segment elevation myocardial infarction diagnosis as described in Ruggeri and Soriano (2014) .

4.4.3. Density Based Silhouette Information (DBS)

In general silhouette analysis is a form of internal cluster validation which can be used to study the separation distance between and within the resulting clusters. It validates the clustering performance based on the pairwise difference of between and within-cluster distances. This index can also be used to determine the optimal cluster number through maximizing the value of this index, Liu et al. (2010). The silhouette plot displays a measure of range $[-1,1]$ which show how close each point in one cluster is to points in the neighbouring clusters and thus provides a way to assess number of clusters visually. Silhouette index (as these values are referred to as) near $+1$ indicate that the sample is far away from the neighboring clusters. A value

of zero indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster. The thickness of the silhouette plot can also give an idea of the cluster size. Silhoutte information can also be used to describe cluster compactness, a description of how objects are closely related in a single cluster although this can be done via variance analysis. Therefore, silhouette information can be used for both within and between cluster diagnosis.

The between and within distances are not always easy to calculate and apply with nonparametric methods. The idea of silhouette information usage in nonparametric density-based clustering procedure was developed by Menardi (2011). The nonparametric dbs was made possible via incorporating the idea of both within cluster distances as well as cluster posterior probabilities. An idea which was used to develop a nonparametric density based clustering algorithm called **PdfCluster** package in R by authors like, Azzalini et al. (2014) and Menardi et al. (2014). The dbs are calculated from the fact that if $x_i \in \chi$ is drawn from a probability density function f, one can evaluate the posterior probability that it belongs to group υ_m, $m = 1 : M$ as:

$$\tau_m(x_i) = \frac{\pi_m f_m(x_i)}{\sum \pi_m f_m(x_i)} \tag{4.12}$$

where π_m is the prior probability of υ_m and f_m is the density of group υ_m at x_i. Then the dbs is defined as follows

$$dbs_i = \frac{log\left(\frac{\tau_{m_0}(x_i)}{\tau_{m_1}(x_i)}\right)}{max_{j=1:n}\left|log\left(\frac{\tau_{m_0}(x_j)}{\tau_{m_1}(x_j)}\right)\right|} \tag{4.13}$$

where m_0 is such that x_i has been classified in υ_{m_0} and m_1 is the group into which τ_m is maximum $m \neq m_0$. Thus a dbs is vector reporting the density-based silhouette information of the clustered data clusters can be formed, Azzalini et al. (2014), Menardi et al. (2014), Menardi (2011). These can be used as a diagnostic tool for cluster quality evaluation as proposed by Menardi

(2011). A positive dbs indicates well classified observations whilst a negative dbs indicates misclassifications. The existence of partitioning within cluster dbs plots indicate the presence of hidden clusters (spurious).

4.4.4. Bandwidth Selection

The selection of an appropriate value of h the bandwidth parameter which affects the density estimations dramatically as explained in section 4.2.3 in nonparametric methods is vital for best cluster results and estimations. The choice of kernel since it is not as influential we are going to choose Gaussian kernels for all the estimations. Bandwidth selection methods have been studied intensively especially in no mixture structures by authors who include Härdle et al. (2004), Scott and Sain (2005). Bandwidth selection in nonparametric mixtures can be a very challenging procedure although some standard ideas can still be incorporated, there seems to be challenges on under or over smoothing as explained by Benaglia et al. (2009a). The bandwidth can calculated based on the block and component so that it is very adaptive in the following way:

$$h_{jl}^{t+1} = 0.9\ min\left\{\sigma_{jl}^{t+1}, \frac{IQR_{jl}^{t+1}}{1.34}\right\}(nC_l\lambda_j^{t+1})^{-1/5} \tag{4.14}$$

where $nC_l\lambda_j^{t+1}$ is the sample size for the lth block and j^{th} component estimate, and σ_{jl}^{t+1} and IQR_{jl}^{t+1} are the weighted standard deviation and empirical interquartile range for the lth block and j^{th} component. These can be calculated from the M step of the EM algorithm and weighted interquartile ranges respectively as explained in Benaglia et al. (2011). The methods allow a component block bandwidth selection which is iteratively adaptive opposed to the same bandwidth selections normally used. This brings flexibility in such a way that different blocks or components are allowed to have different properties. Moreover, since the bandwidth is iterative, the estimations can be done prior to knowledge of the mixture structure.

4.4.5. Motivation

Nonparametric mixtures differs diversely from parametric mixtures and the main differences include that: whilst the parametric setting of equation (2.1) clusters are associated with the components f_j, in the nonparametric context model (1.5), clusters are associated with regions with high density as density functions are solely by estimation. Although they may sometimes lead to the same results, the two approaches are very different and it is our ultimate purpose in this chapter to dwell on the nonparametric. This is a more robust way of identifying and inferring subpopulations as it uses modality inference on the number of modalities as alluded by Cheng and Ray (2014) over inference in a finite mixture model (2.1) and hierarchical approach of estimating the number of cluster output. The use of multivariate modal inference using the (KDE) which is normally used in both multivariate and univariate cases. Smooth polynomials can also be used to model the different cluster density estimations as shown by Rebafka and Roueff (2015).

Unlike most researchers who used the kernel density estimation in the nonparametric finite mixtures, Rebafka and Roueff (2015) implemented a different approach via Legendre polynomials as well as Gamma and Beta mixtures. However, in this chapter we will concentrate on nonparametric clustering on multivariate mixed data and in particular kernel density estimators will be used to cluster the data. Although most nonparametric models have been built under the assumption of identical independence of components for identifiability, we use examples of a latent variable model as a mixture of Gaussian distributions. This can be from observed data of mixed type which can be ordinary, binary, continuous etc which then can be estimated and clustered by nonparametric models as shown in McParland et al (2016). A nonparametric model based clustering in recent works has also shown that the conditional independence assumption can be dropped at times and adopt the independent component analysis as described by Zhu et al. (2015).

4.5. Nonparametric Mixture Models Applications

4.5.1. Data

The data used in this work is found in the UCI database for the Cleveland Clinic Foundation for heart disease as given by Detrano. Original data has about 76 raw attributes and 303 observations for each attribute. In this research only 14 attributes were used and they are described as in table 0.1. of appendix. Variable distributions are given as figure 0.1., 0.2., 0.3. of the appendix.

4.5.2. Packages Used

The Cleveland heart disease data is a mixed data type. Therefore, in order to cluster it using nonparametric mixture models via graph theory methods there was need to measure it dissimilarities among the points. This was done using the Gower [Gower and Ross (1969)] coefficient in the **cluster** package in R. The cluster package also enabled us to classify multidimensional scaling of the data under the ***cmdscale*** function as described by Maechler et al. (2016). Nonparametric density approximations are then implemented on the mixed heart disease data using a Gaussian KDE. These are applied and implemented through the **PdfCluster** package by Azzalini et al. (2014). The PdfCluster automatically selects the procedure to be used for detecting connected components of the density level sets, depending on the data dimensionality. This is enabled by making an internal call to function **kepdf**() both to estimate the density underlying the data and to build the connection network when the pairwise connection criterion is selected.

A kernel density estimation with Gaussian kernel is built, with the vector of smoothing parameters set to the one asymptotically optimal under the assumption of multivariate normality. We used both a fixed shrinked smoothing parameter and adaptive bandwidth since the data is highly multidimensional. Cluster diagnosis in this method uses the **dbs** criteria in the pdfCluster package as described before in section 4.4.3. The number of groups here will be identified as the number of modes of the estimated

density used. Analysis is to be done with different number of clusters and diagnosed using the dbs criteria.

4.6. Using KDE with Adaptive Bandwidth

In this section nonparametric miture models were analysed using a Gaussian kernel with a adaptive bandwidth. varying principal components will be analysed as well. Figure 4.3. shows how different principal components have different clusters and hence results in different dbs median values for each cluster. Cluster information for a 2 principal component model (i.e $p = 2$) is shown as figure 4.3.(a). It shows that a 2 component model has 2 clusters both with positive dbs median values and without any cluster partitions. This is an indication that the clusters are both well classified and separated. Their high positive dbs median values indicates well separated clusters whilst absence of partitioning with cluster dbs shows cluster do not overlap. The quality of clusters seems to decrease as the number of principal component increases to 3 as shown by figure 4.3.(b).

It shows the existence of spurious clusters due to presence of partitions within a cluster dbs diagram, an indication that clusters are not well separated. When $p = 3$, a total of 5 clusters with consists of three spurious clusters shown by presence of partitions within cluster and 2 negative dbs valued clusters. Negative dbs values explains that most of the observations classified under a given cluster were wrongly classified. Therefore the 3 component model has both overlapping clusters and well as wrongly classified observations which makes it a bad model.

Figure 4.4.(a) and (b) represents the dbs cluster information for clusters formed with 4 and 5 principal components respectively. Whilst the 4 component model has 5 clusters, the 5 component has 4 clusters most of them shows partitions and with negative dbs values. Of the 5 clusters from figure 4.4.(a), 3 are spurious whilst 2 have negative dbs median values a indication of high poor quality clusters. The same trend is shown in 5 component model in figure 4.4.(b) which has a combination of both ill separated

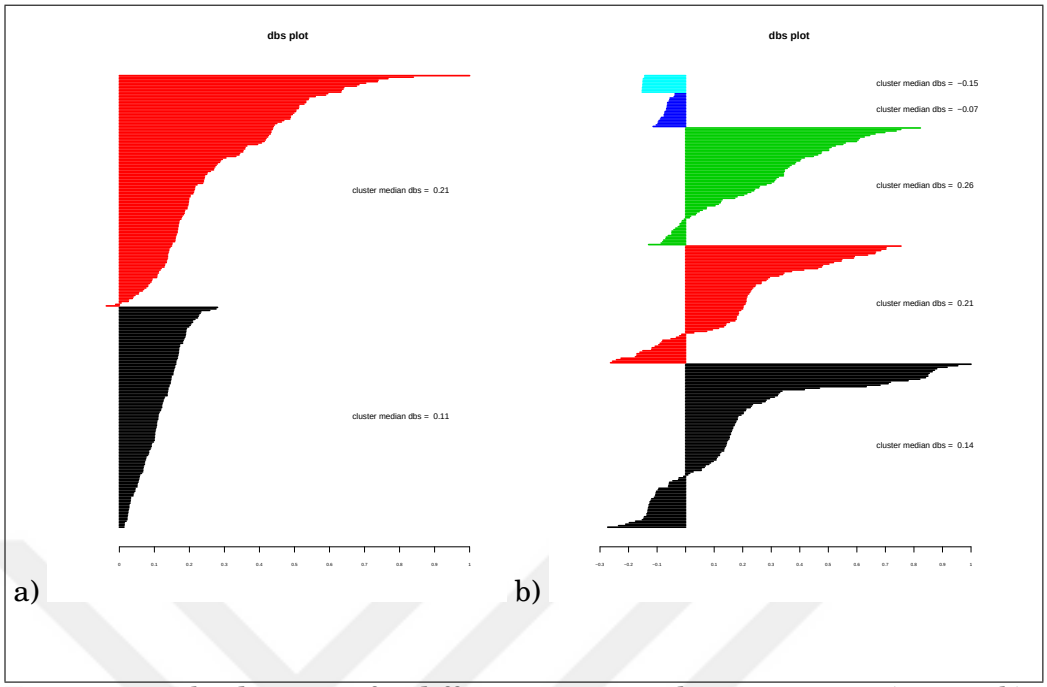


Figure 4.3.: Dbs diagrams for different p principal components: a) $p = 2$, b) $p = 3$

and ill classified clusters.

Figure 4.5. shows the dbs mean values for clusters formed with 8 and 10 principal components. There are 5 clusters from a 8 component model whilst there are only 2 with a ten component model. However, although the 10 principal component model gives the same number of clusters with the 2 principal component model, both clusters have negative dbs values and one of them is partitioned a high indication that observations in those clusters were misclassified and are highly ill partitioned.

Therefore given the dbs values for each cluster we conclude that data fits well in a 2 component model with 2 clusters. In this case the best model has the number of principal component corresponds to the number of clusters although it should be noted that different results can be found in other situations. We are therefore safe to classify individuals exposed to heart disease risks into 2 categories depending on which risks are they exposed to at a given time. This makes sense in the fact that the nonparametric model is able to cluster those individuals who are likely to get heart disease early in

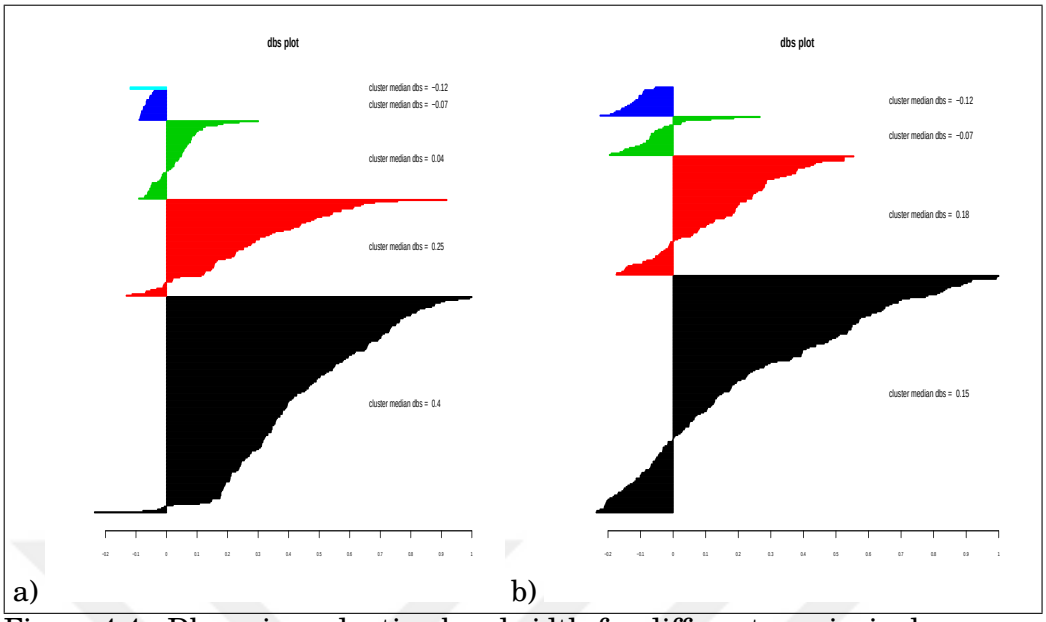


Figure 4.4.: Dbs using adaptive bandwidth for different p principal components: a) $p = 4$, b) $p = 5$.

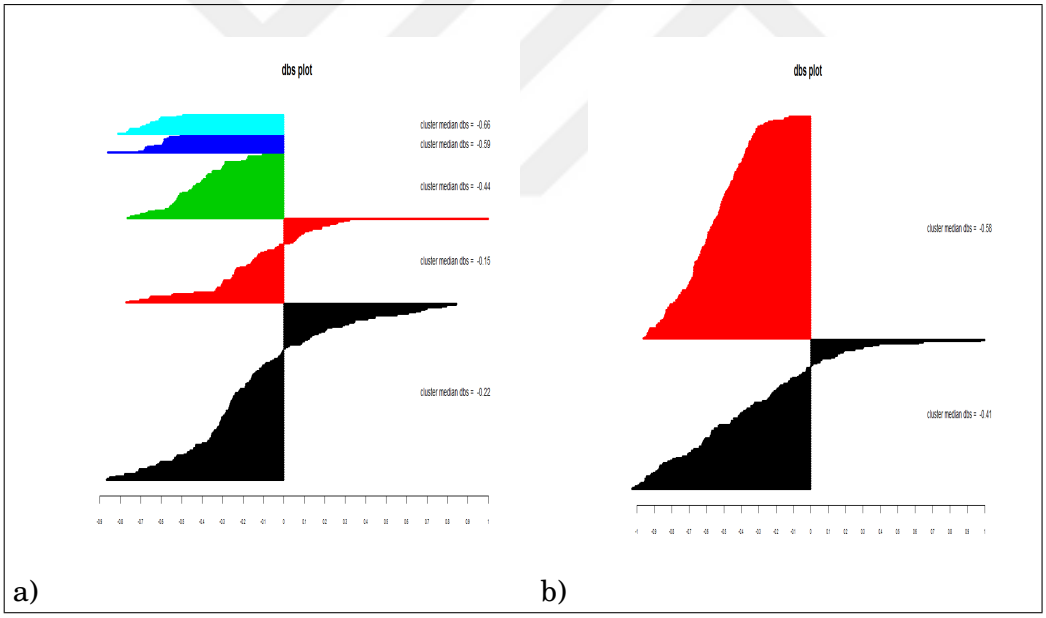


Figure 4.5.: Dbs diagrams for different p principal components: a) $p = 8$, b) $p = 10$.

one group and those unlikely to in a different cluster the same result produced by Mufudza and Erol (2016) with a parametric mixture models.

4.6.1. Cluster Scatter Plots under Adaptive Bandwidth

The cluster scatter plots represent all the observations in the two components categorised by their clusters. It can help to view cluster partitioning besides the use of dbs plot. The cluster scatter plots for some of the models is thereby shown in figure 4.6.: where the cluster scatter plot for a

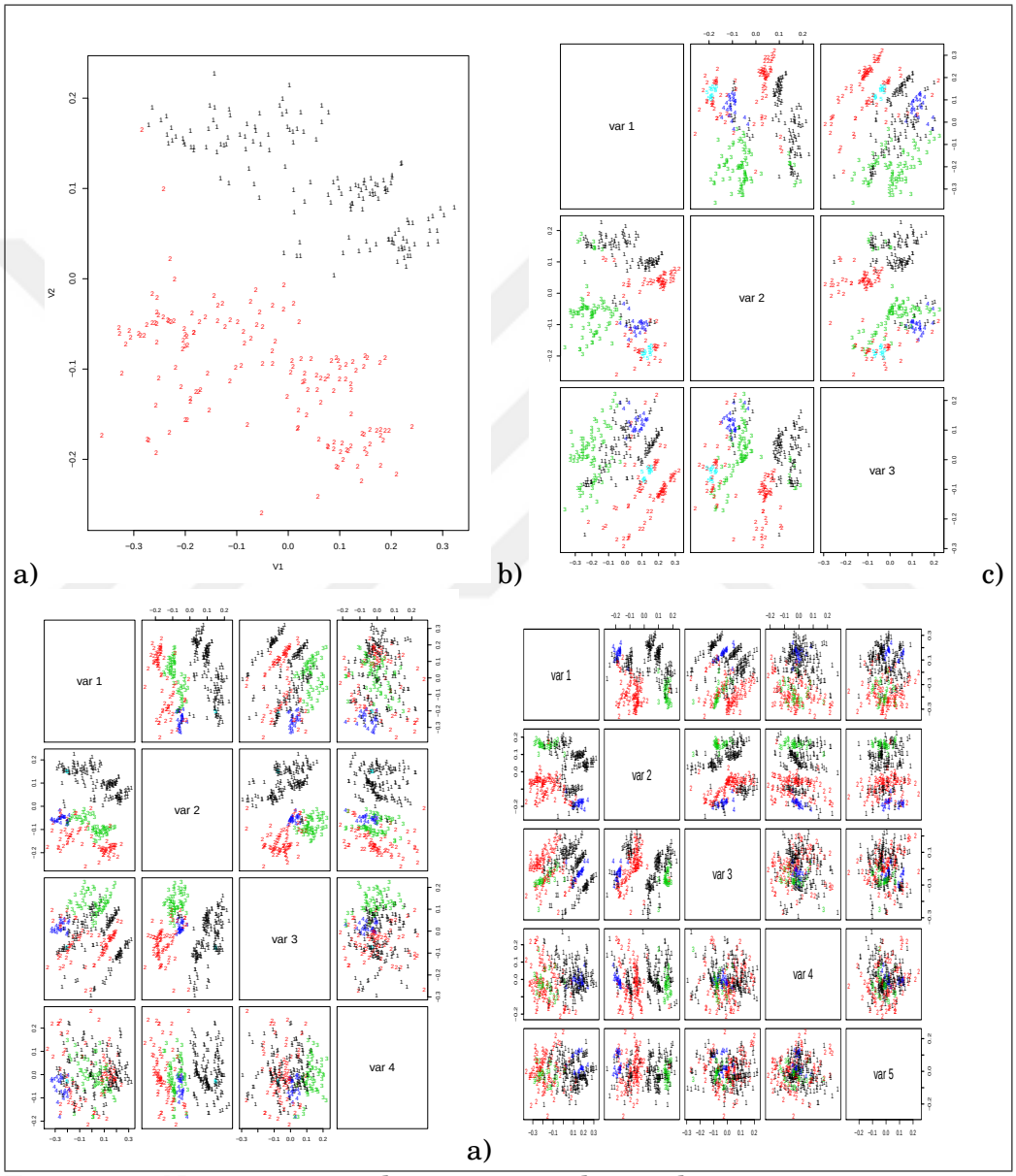


Figure 4.6.: Nonparametric cluster scatter plots with 2 components

2 and 3 component models are shown as figure 4.6.(a) and (b), respectively. It shows that well separated clusters are produced by a 2 component as in

figure 4.6.(a). However, decrease in clarity between the clusters seems to increase as the number of component increases as evidenced by a 3, 4 and 5 component models in figures 4.6.(b), (c) and (d), respectively. They show that it is very difficult to separate the points from within different clusters, an indication that there is overlapping on the clusters as a result of ill separation and wrong classifications. Thus we choose a 2 component 2 cluster model as the best model even using cluster scatter plots.

4.6.2. Cluster Summary Properties under Adaptive Bandwidth

Considering all the models analysed under varying principal components we discovered that different components produced varying number of clusters. A summarized table to show the model properties of the different models is given in as table 4.2.. It generally, shows that when adaptive band-

Table 4.2.: Cluster Summary Under Adaptive Bandwidth

Comp.	Total Clusters	Spurious Clusters	-ve dbs Clusters
2	**2**	**0**	**0**
3	5	3	2
4	5	3	2
5	4	3	2
6	5	3	4
7	**2**	2	**0**
8	5	2	3
9	5	2	5
10	**2**	2	2

width is used, an increase in the number of components have a negative impact on cluster quality as it seems to decrease cluster quality. The decrease in quality is evidenced by increase in the number of spurious clusters and presence of negative dbs median valued clusters. Thus more wrong classifications with ill partitioning comes with increase in the number of components. Hence, the best clusters are produced by 2 component model as they have neither partitions nor negative dbs values.

It should be noted that although 7 and 10 component model produced the same number of clusters as 2, both clusters from a 7 component model

Table 4.3.: 2-2 Model DBS Cluster Summary

Cluster	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	0.01311	0.06579	0.11020	0.11450	0.16130	0.28030
2	-0.03721	0.14310	**0.21280**	**0.28590**	0.43170	1.00000
Data	-0.03721	0.09534	0.15800	0.20220	0.23520	1.00000

are spurious and although they have positive median dbs the values are very low below data mean dbs values compared to a 2 component model. This indicates poor cluster separation and relatively high miclassifications on the clusters. In the same manner the 10 component model have all negative dbs values for the 2 clusters a high indication that clusters were wrongly classified. We therefore conclude that individual under risks of heart disease be classified by 2 component nonparametric density model with 2 clusters.

4.6.3. Two Component Two Cluster Model Properties

In this section we are going to infer on the 2 cluster 2 component model chosen as the best model when adaptive bandwidth is used. This involves getting the properties of each of the clusters produced by the Pdf-Cluster algorithm. Some of the model properties for each cluster including statistical measures of the clusters are given as table 4.3.: It shows that cluster 2 has high dbs median and mean values of 0.21 and 0.29 respectively compared to cluster one which is even higher than the whole sample data. It also has a high max positive dbs of 1 indicating quality clustered data. The width of the dbs graphs for the 2 component 2 cluster model are also reasonably bigger than any other

4.6.4. Surface Densities for the 2-2 Model

The surface densities for the nonparametrically clustered data with 2 clusters and 2 components is shown as figure 4.7.. The mixture density can be represented by 2 clusters where the first one has develop heart disease at very low probability whilst the second one has very high probabilities to develop heart disease. The variable densities under the 2 component 2 clusters model groupings using nonparamteric mixtures is given by fig-

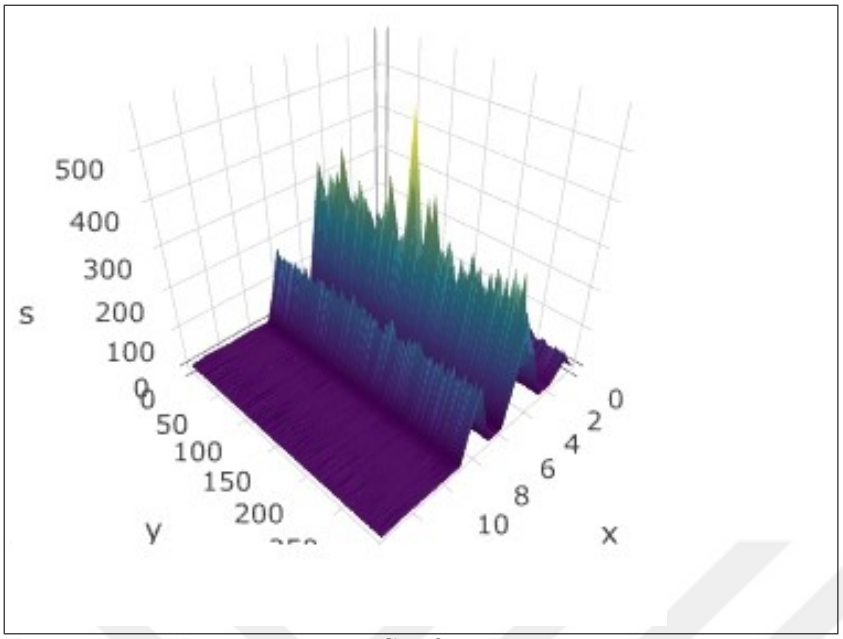


Figure 4.7.: Nonparametric Surface Densities

ure 4.8. where (a)-(h) represent age, sex, cholesterol, resting blood pressure, maximum heart rate, resting electrocardiographic , chest pain and defects involved, respectively. Other variables densities are given as figure 0.4. in appendix.

4.7. Nonparametric Mixtures with Fixed Bandwidth

In this section we are going to use and analyse the same mixture model with a Gaussian kernel nonparametrically. However, a fixed bandwidth will be used for all the models with varying components and clusters as explained in section 4.6. Figure 4.9. shows the dbs plot of the nonparametric mixture model with 3, 4, 7 and 10 components, from (a)-(d) respectively. The 3 component model has 3 clusters which has all have positive dbs median values above the mean dbs for the whole data. They are also reasonably thick an indication that most observations were clustered in those clusters with the thickest cluster having a median dbs value of 0.44 whilst the thinnest has dbs of 0.10. The 4 component model also has 3 clusters although two of them show the existence of hidden clusters. Although all dbs median values are positive, most of them have dbs median values that are

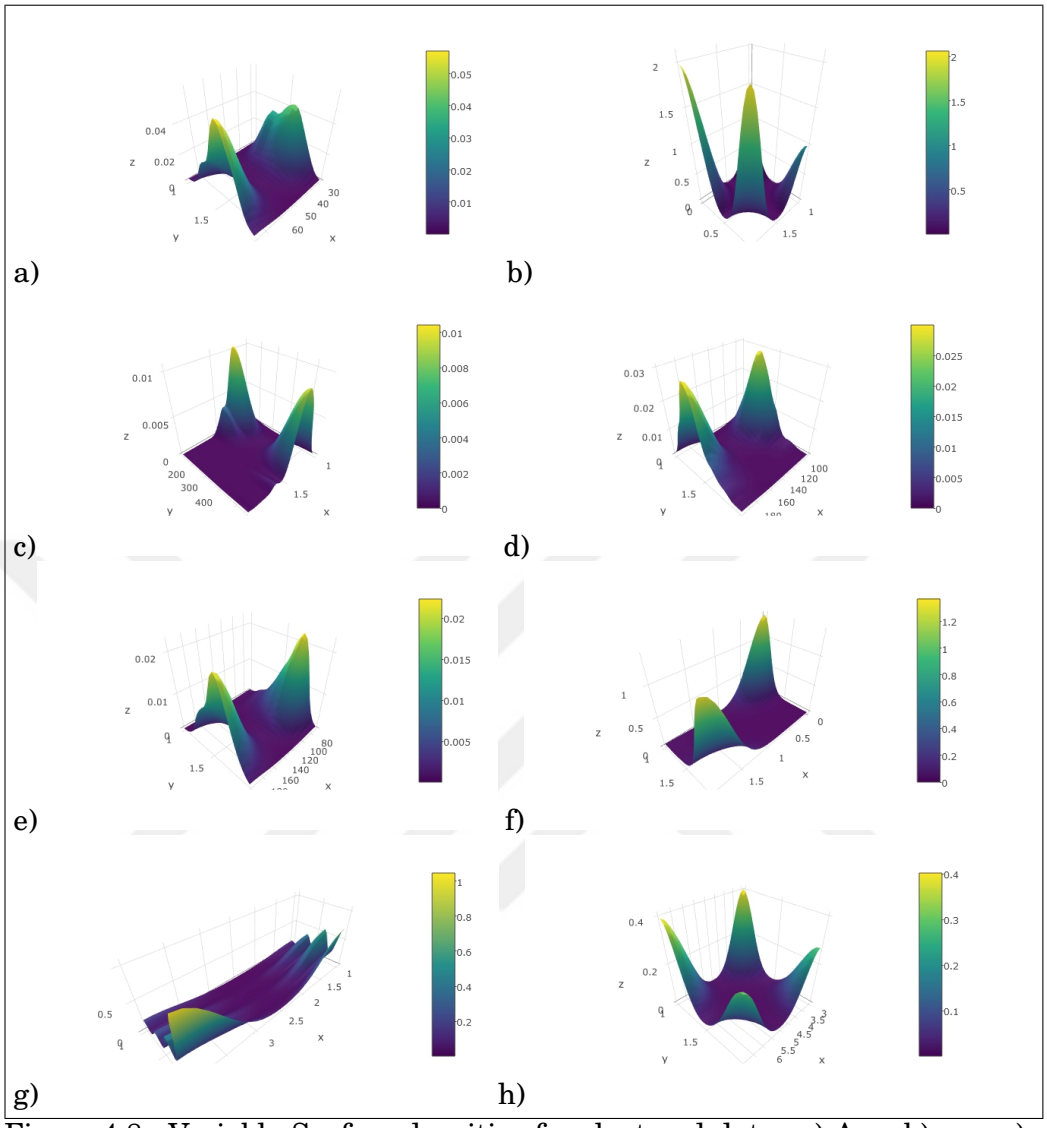


Figure 4.8.: Variable Surface densities for clustered data: a) Age, b) sex, c) cholesterol, d) resting blood pressure, e) maximum heart rate, f) resting electrocardiographic, g) chest pain, h) defects

way too low than the mean dbs value for the data. This indicates that some observations may have been wrongly clustered hence it is very difficult to distinguish the clusters clearly. The results also show that there is an increase in the number of clusters as the number of components increase. However, the quality of the clusters decrease as indicated by the increase in the number of clusters with hidden (spurious) clusters and existence of either very low dbs mean values or negative dbs values. Figure 4.9.(c) and (d) shows the

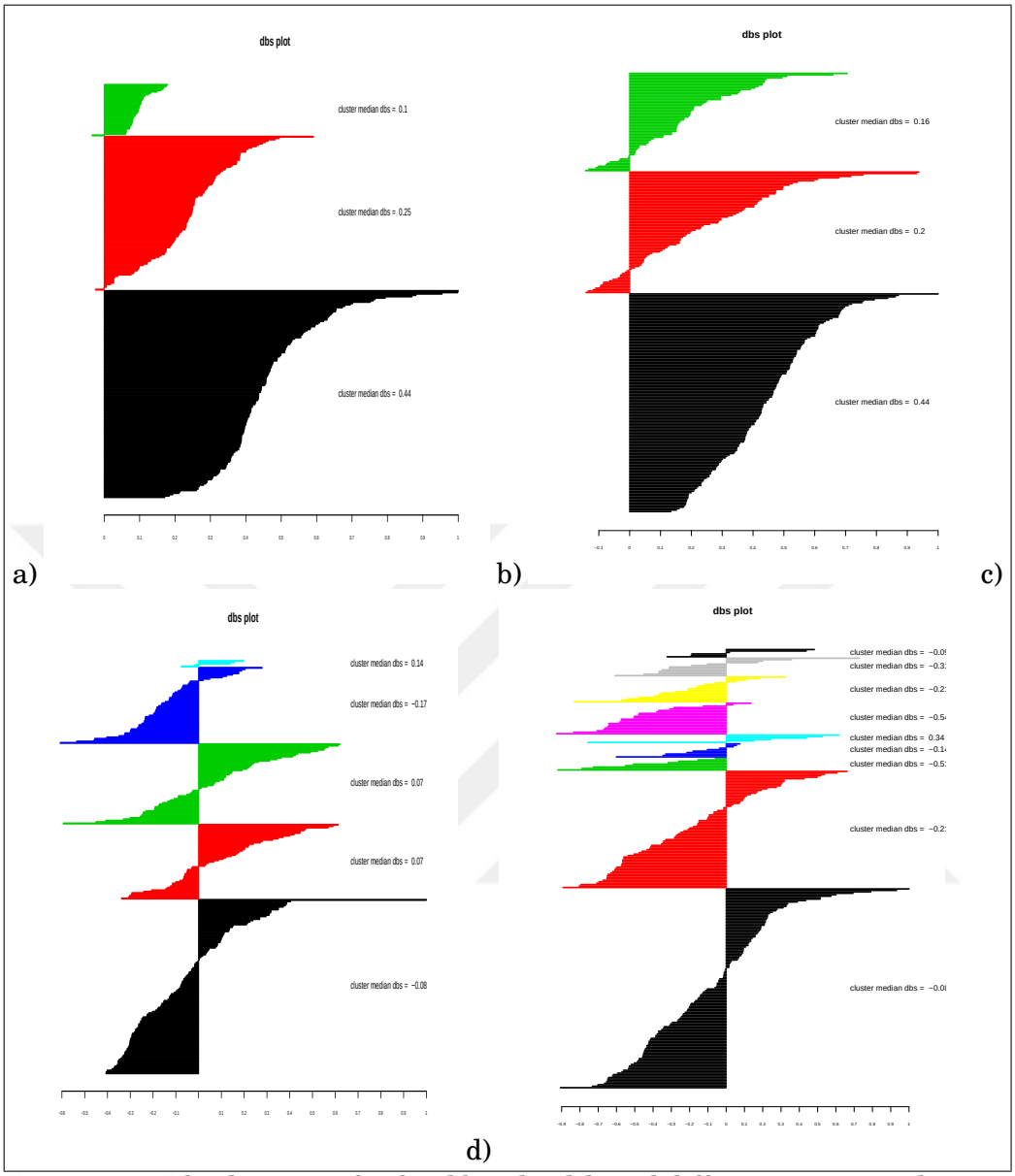


Figure 4.9.: Dbs diagrams for fixed bandwidth and different p principal components: a) $p = 3$, b) $p = 4$

properties of the clusters produced when the components increases to 7 and 10 components, respectively. A total of 5 all spurious clusters are produced with 7 components showing that the partitioning is poor with two negative dbs valued. In figure 4.9.(d), a ten component model with 9 clusters which are highly spurious and 8 negative dbs values is a proof of poorly classified and partitioned clusters.

4.7.1. Cluster Scatter Plots for Fixed Bandwidth Models

Cluster scatter plots for a 2, 3, 7 and 10 component models are shown as figure 4.10.(a)-(d), respectively. It shows that both the 2 and 3 component models has 3 clusters. However, we have well separated clusters on a 2 component model opposed to 3 component one. The same is seen on a 7 component model which also has 5 interdependent clusters whilst 10 component model is an ugly model where clusters cannot even be recognised. Whilst partitions in a 2 component 3 cluster model can be easily seen, it is not the same as we increase number of components as shown in figure 4.10.(b)-(d). It is difficult to distinguish clusters for observations especially in component 1 and 3 as well as observation between clusters 2 and 3 as there seems to be cluster overlapping for the 3 component model. The models with varying components were not shown. However, they are too congested and hence clusters in all the components cannot be distinguished properly. Thus we choose a 2 component 3 cluster model when fixed bandwidth is used.

4.7.2. Summary: Fixed Bandwidth Models

A summary of the analysis on the different models with varying components with different clusters when a fixed bandwidth is employed on the Gaussian kernel is given in table 4.4.. It shows that the number of clusters slightly increase with increase in the number of components. However, the quality of the clusters is not desirable as there seems to exist a lot of spurious clusters although classification seem right due to positive dbs for clusters up to 6 and gets worse as we increase the components. Although there are no clusters with negative dbs median values for 3-6 component models, the majority of the clusters have poor partitioning as indicated by very low dbs values which are close to zero and way below mean dbs value for the data. Given the summary a 2 component 3 cluster model has the best clusters when the bandwidth is fixed.

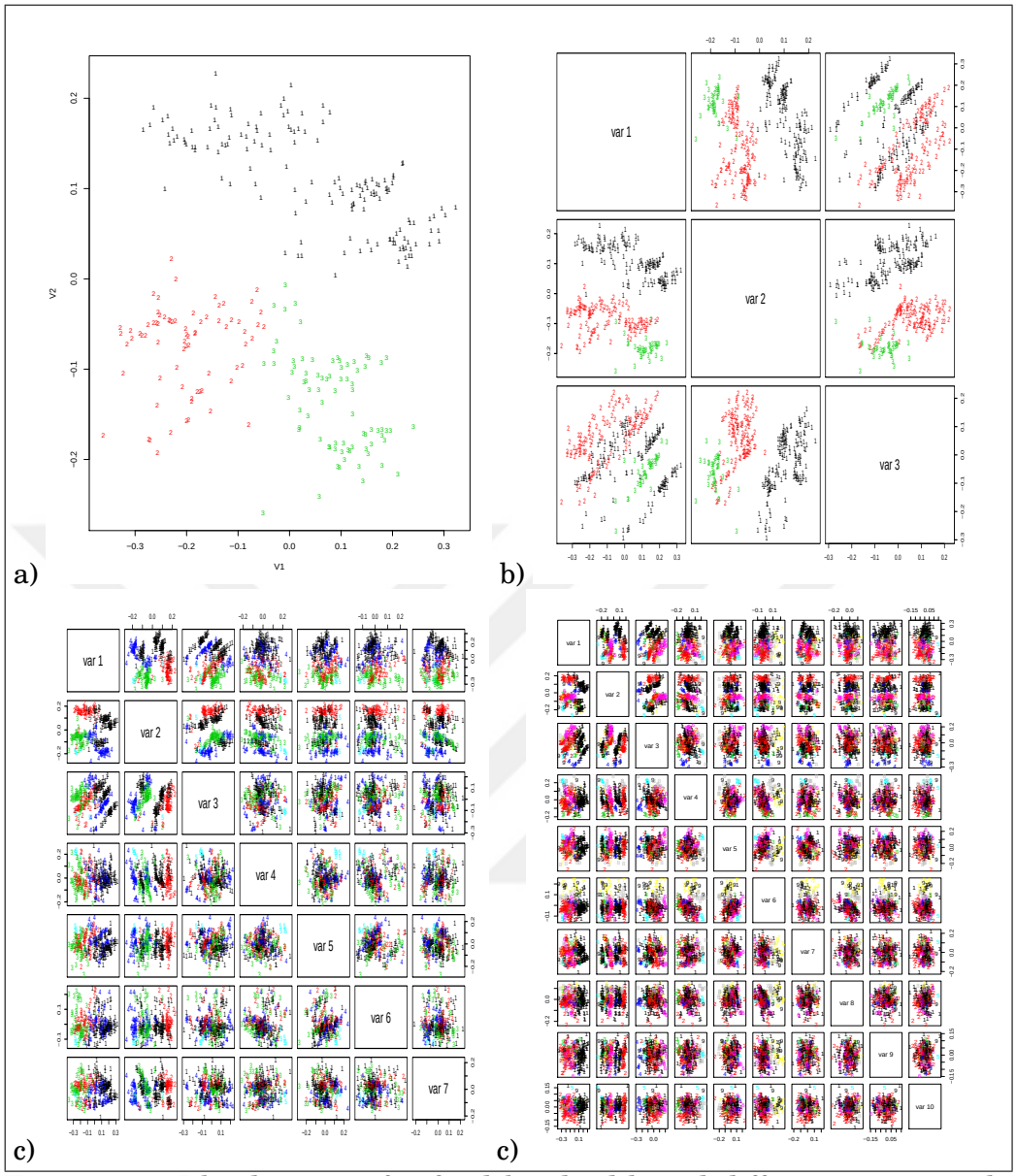


Figure 4.10.: Dbs diagrams for fixed bandwidth and different p principal components: a) $p = 3$, b) $p = 4$

4.7.3. Fixed Bandwidth Best Model Properties

The properties of the 2 component 3 cluster model which proved to be the best model for fixed bandwidth is given in table 4.5.. The mean dbs for the first cluster is greater than the overall mean dbs for the data with the other two cluster slightly lower than the mean dbs although they are still greater than the 25 percentile.

Table 4.4.: Cluster Summary for Models with Fixed Bandwidth

Comp.	Total Clusters	SpuriousClusters	-ve dbs Clusters
2	3	na	na
3	3	2	0
4	3	2	0
5	4	4	0
6	4	3	0,(**NB** $\exists$ 0)
7	5	5	2
8	8	7	5
9	9	9	7
10	9	8	8

Table 4.5.: Best Model properties with Fixed Bandwidth

```
Density based silhouette summary of cluster
1 :
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max
0.007891 0.261700 0.374800 0.375500 0.476900 0.86500

2 :
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max
-0.02515  0.10100  0.21780  0.21660  0.27320  1.0000

3 :
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max
-0.02658  0.08773  0.19750  0.22140  0.30080  0.6756

Density based silhouette summary of data
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max
-0.02658  0.17890  0.27390  0.29650  0.42260  1.0000
```

4.8. Conditionally Independent Densities for Nonparametric mixtures: MvnpEM

This is appropriate for model based clustering using nonparametric mixture methods when we assume conditionally independent component densities of multivariate data as described in Chauveau (2015), Young et al. (2016) and many more. The multivariate nonparametric clusters are determined via nonparametric method and analysed under EM like algorithm as described by Benaglia et al. (2009a) and Benaglia et al. (2009b) as explained in section 4.3.1 and 4.3.2 which are basic derivations from the nonparametric equation model (1.5).

Data:

The application was done on the Cleveland heart disease data described in

Detrano. The data properties has been explored in section 6.6 and 3.3.3 from the appendix.

Packages

We use the **mixtools** package in R specifically the multivariate nonparametric clustering function **mvnpEM** as described by Benaglia et al. (2009b). The analysis result was administered for different models and with different components and blocks.

4.8.1. Model Results

Analysis was done for different componens and blocks in R. Then best models was choosen based on the integrated square errors. Basing on the square errors which represent the variance matrices, The best models had 3 blocks and 2 component. The number of components agrees with the principal component analysis done to the data which also had 2 components as the best model. All the analysis were done using mvnpEM in R calculating the means and standard deviations each block and component. A 2 component 3 block was choosen as best model basing on te analysis done and the model properties are given as table 6.6 in appendix. The first block has coordinates age, resting blood pressure, cholesterol levels,maximum rate and ST depression due to exercise. The mean and the standard deviations are given componentwise as in table **C1** and table **C2** for all the 3 blocks as shown in table 6.6. Second block has 4 coordinates which include sex, fasting blood sugar, exercise induced angina and number of diagnosis whilst the last block has all the other remaining coordinates. The conditional independence assumption then assumes that the blocks are independent. The posterior probabilities for each of the components are given by:

$$\lambda_1 = 0.3847056,\ \lambda_2 = 0.6152944.$$

The bandwidth

$$\mathbf{B} = \begin{bmatrix} 2.5945267 \\ 0.1341370 \\ 0.2142147 \\ 4.2842932 \\ 13.7097381 \\ 0.1022458 \\ 0.2856042 \\ 6.5662158 \\ 0.1348534 \\ 0.3332838 \\ 0.1768863 \\ 0.2142147 \\ 0.5564082 \\ 0.3526483 \end{bmatrix}$$

The integrated square mean errors for the components and blocks are given in table 4.6. where component 1 and seems to be more compact than component 2 due to is low integrated error values. Block 2 is the most compacted of all the blocks in both the components.

Table 4.6.: Integrated Square Errors

	Blocks		
	1	2	3
Components			
1	0.000210011	5.121460e-06	0.0001309104
2	0.000409646	1.941282e-06	0.0001151588

4.8.2. Component Integrated Error Densities

The density distributions for the integrated mean square errors for each of the component and block is shown by figure 4.11.. The f_{ij} represent the density function for the ith component and j^{th} block for $i = 1,2$ and $j = 1,2,3$. It shows that the errors turn to zero for most of the blocks and components except component 2 and block 2 where it is significant. An indication that block 1 of component 2 is less compact compared to all other blocks in the different components.

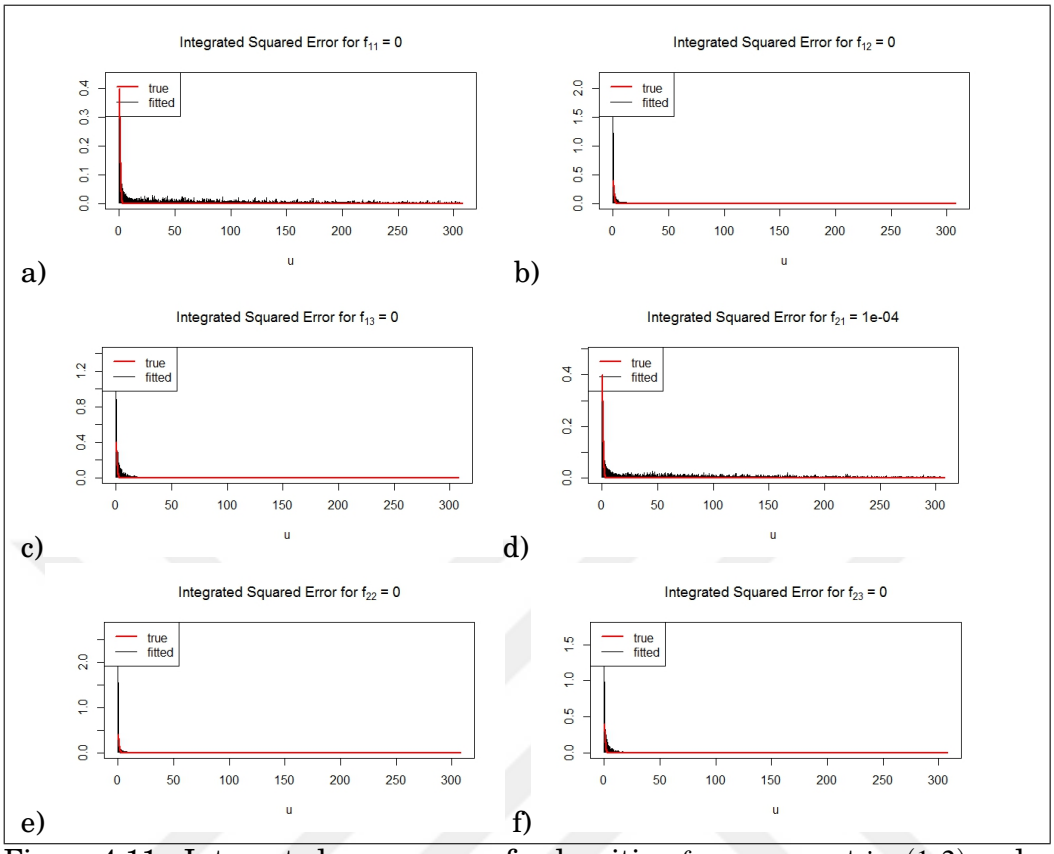


Figure 4.11.: Integrated square error for densities f_{ij} component $i = (1,2)$ and block j $j = 1,2,3$.

4.8.3. Clusters from MvnpEM

The clusters for the models with different components under 3 blocks are shown in figure 4.12.. The quality of cluster for models with 3 blocks and varying components can be shown via the cluster distribution as shown in figure 4.12.. Note that analysis was also performed under different number of blocks not shown here, however, the 3 block model was a better model. Under the 3 block models, figure 4.12. shows the cluster distribution for models with 3 blocks and for components ranging from 2-5 as represented by figure 4.12.(a)-(d), respectively. The 3 block model with 2 components and 2 clusters has well separated clusters than all other models. This indicates good quality clusters were clusters are well separated and distinguished with rightly classified observations. The quality however, seem to decreases as components increase. The 3 block 3 component model has 3 clusters, 2 of which

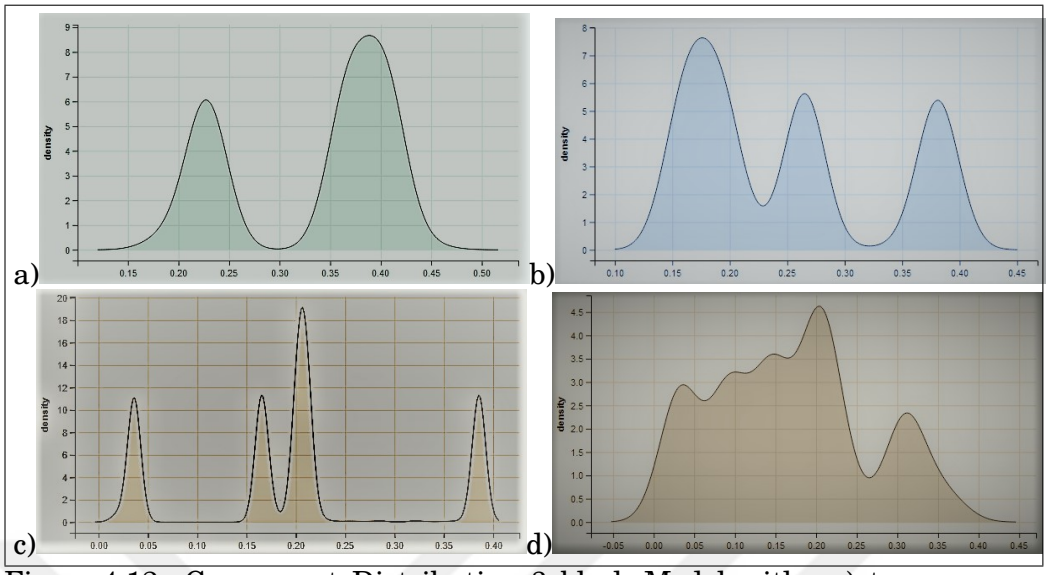


Figure 4.12.: Component Distribution 3 block Model with: a) two components, b) 3 components, c) 4 components and d) 5 components

are overlapping and the same trend seen with a 4 and 5 component models. Thus we choose a 2 component model with 3 blocks as the better model for the data under the conditional independence assumption.

4.8.4. Componential Cumulative Distribution functions

The cumulative density function for the different models with 3 blocks were estimated for models with components 2, 3 ,5 and 7 respectively as shown in figure 4.13.. This enables us to see the risk involved with each of the components where components are high risk and low risk. The risk of individuals in both components seems to be above 70%. In figure 4.13.(a) the distribution shows that more risk is associated with individuals in cluster 2 compared to those in cluster 1 given a 2 component model. However, there is a slight increase in the percentage risk as risk increases in cluster 2. There is evidence that indeed heart disease diagnosis cannot be generalised under the same blanket term of risks as different groups progresses at different risk rates as given by figure 4.13.. Thus different clusters has different level of risk depending on which risk is more evident than the other. An interesting point to note is how nonparametric mixture models differentiate the heterogeneity among observations by grouping them into clusters of relatively

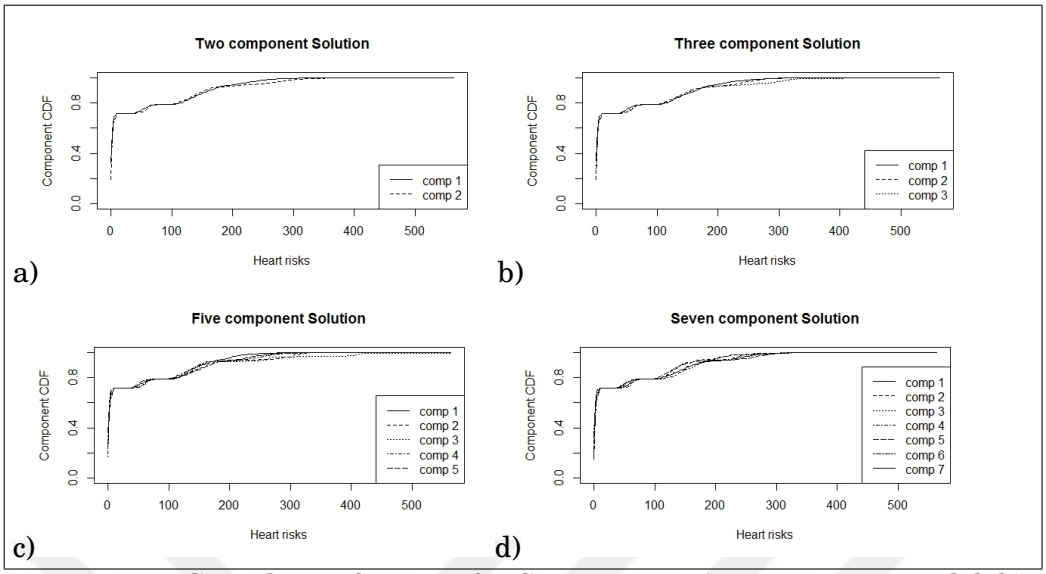


Figure 4.13.: Cumulative function for Components:a) 2 component model, b) 3 component model, c) 5 component model, d) 7 Component model

similar qualities. In this thesis we observed that given the heart disease risks which are variables in this case, indeed heart disease best described in terms of 2 clusters where one has high risk with the other having a lower risk. Since the 2 component model was chosen, it was noted that individuals in cluster 2 were at more risk to develop heart disease compared to those in cluster 1. The risk for other models were quite difficult to distinguish a factor which may be facilitated by the fact that clusters are difficult to differentiate. Therefore, we can safely deduce that say the best model classifies individuals into high risk and low risk levels.

5. DATA LABELING VIA MIXTURE MODELS

5.1. Introduction

Data labeling typically takes a set of unlabeled data (daa with unknown characteristics) and augments each piece of the unlabeled data with some sort of meaningful tag, label, or class that is somehow informative or desirable to know. For example, labels for the heart disease patients can be whether an individual is on high risk or low risk. Labels for data are often obtained by asking humans to make judgments about a given piece of unlabeled data. However, this is significantly more expensive to obtain than the raw unlabeled data. Thus, cheaper methods to label data, such as clustering can be used for data labeling. Clustering can be done as a pre process in a classification problem. This is normally a helpful labeling technique especially in the case where the outcome classes are binary. The obtained groups can then be used as labels for an unlabeled data and then classification can be administered.

The clustering can be done parametrically or nonparametrically via different methods. Thus, heterogeneity can then be explained in two fold using both clustering and classification. Example by López et. al. (2012) demonstrated how clustering can be used as labeling techniques for classification in determining students marks to determine failure or pass status for students. They used different clustering methods to identify if student participation in the course forum can be a good predictor of the final marks for the course and to examine whether the proposed classification via clustering approach can obtain similar accuracy to traditional classification algorithms. A similar study in predicting student performance on online forums was predicted by Romero et al (2013) using classification enhanced by clustering.

Data labeling using clustering techniques can therefore, be considered as a semi supervised classification method. This is due to fact that it involves a combination of supervised classification via clustering as well

as supervised classification via the different methods of classifiers to allow group prediction. Hence the combination can be semi supervised. This can also work in cases where the part of the data has missing labels. Authors like Sakai et al (2016) alluded that if label information is directly extracted from unlabeled data, then accuracy can be improved. They explained this on how Positive Unlabelled classification (POC) can be incorporated on negative data and proposed a novel semi-supervised classification approach for the system.

In this chapter we will concentrate on clustering via mixture models. Both parametric and nonparametric mixture models are going to be examined as a labeling technique to classification. As explained before in chapters 3 and 4, the labels will be taken from the clusters or groups identified in the heart disease data. Application data will be same heart disease data by Detrano to determine the possible classes involved in heart disease diagnosis. Class label will be the groups obtained on our previous analysis in chapter 3 for parametric mixture models, whilst groups obtained in chapter 4 will be used for nonparametric mixture models. Quality of the labels and evaluation on the different mixture models will be then administered over the traditional classification algorithms.

5.2. Classification

Classification is a data mining technique that allows you to use your past data instances, with known output values (labeled), to predict an unknown output value of a new data instance. It is known as a supervised machine learning technique in general terms whilst clustering is refereed as unsupervised machine learning technique. Supervised classification involves so many algorithms which can be tested on the then labeled data to predict new set of data or compare different classifiers. These can be trees, rules, functional, meta a way which describes the way classifiers classify the data into the labeled classes. Thus, classification is a model based supervised techniques used to predict class of an observation based on some historical

characteristics and unknown categories. The class labels in our case comes from clustering outcomes obtained either via nonparametric mixture models or parametric mixture models which are then used to predict new data. Therefore, whilst clustering is an unsupervised way of grouping data using a measure of similarity, classiïňĄcation is the supervised version which seeks to use the labeled data to improve or classify new data.

5.2.1. Classifiers

In this thesis, a selective number of classifiers were used to classify using the class labels from both nonparametric mixtures and parametric mixture models. In particular with heart disease data, identification of the best classifier for the heart disease prediction will be considered. Questions like, do we have to use parametric or nonparametric clustering to label data? Which method is a better method and why? In particular for heart disease diagnosis, how do we ought to label an individual under high risk or low risk? will be addressed. That is, the goal is to label efficiently and develop a rule, a program, or a procedure that classifies new examples (in the test set) by analysing examples which already have a class label. Six classifiers for heart disease prediction were considered, these include:

5.2.2. Naive Bayes

Naive Bayes implements the probabilistic Naïve Bayes classifier. This method goes by the name of Naïve Bayes because is is based on Bayes' rule and "naïvely" assumes independence. The independence attribute is a simplistic assumption (given the class) in real life data. But despite the disparaging name, Naïve Bayes works very effectively when tested on actual datasets, particularly when combined with some of the attribute selection procedures according to Witten and Tibshirani (2011). The classifier uses Bayes theorem given class variable, y and dependent feature vector $\mathbf{x}_1$

through to $\mathbf{x}_n$, Bayes theorem states the following relationship

$$P(y|x_1,.....x_n) = \frac{P(y)P(x_1,.....x_n|y)}{P(x_1,.....,x_n)}$$

where $P(x_i|y,x_{i-1},x_{i+1},.....x_n) = P(x_i|y)$ by the independence assumption. Thus since the $P(y|x_1,.....x_n)$ is constant, we can use the classification rule as

$$P(y|x_1,.....x_n) \propto \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1,.....x_n)} \tag{5.1}$$

5.2.3. Logistic

Suppose the approximate probability of a system S_p is $P(y_q|S_p,x_q)$, corresponding to a certain input x_q and categorical output y_q. Inorder to explore the system we assume knowledge from a previous observation system behaviour. Since y is categorical this becomes a classification problem and using a logistic function to classify result in logistic regression classification given by

$$\pi_q = \frac{1}{1+\exp(-(1,x_q^T)\beta)}$$

where x_q is the input vector, β is the parameter vector and π_q is probability of y_q to be in I

$$\pi_q \equiv P(y_q = 1|S_p,x_q) \tag{5.2}$$

5.2.4. IBK

It takes the distance between the unknown data point and every known data point needs to be computed. Finding the distance is really quite trivial with a spreadsheet, and a high-powered computer can zip through these calculations nearly instantly. Use the distance to predict the outcomes.

5.2.5. Random Forest

This targets classification taking on values $0,1,.....,k-1$ for mode m representing region R_m with N_m observations. Let the proportion of class k

observation in node m be

$$p_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

The common measure of impurity are the Gini

$$H(X_m) = \sum_k p_{mk}(1 - p_{mk})$$

Cross Enthropy

$$H(X_m) = -\sum_k p_{mk} log(p_{mk})$$

Misclassification

$$H(X_m) = 1 - max(p_{mk})$$

5.3. Classification via Nonparametric Mixture Models

Nonparametric mixture models just like any other clustering technique can be used for labeling data classes before classification. Although parametric mixture models can be more informative and easy too construct, the flexibility in nonparametric mixture models brought by freeing the clusters from a fixed given shape and density function can never be underestimated. Flexibility also allows incorporation of the disparity between clusters and components normally assumed the same in parametric mixture models. Thus this section focus on using nonparametric mixture models as a labeling tool for supervised classification for unlabeled heart disease data.

Parametric models can still be used in data labeling, however, most of them are restrictive in shape and properties. For example, the commonly used Gaussian model is always uni-modal and symmetric a situation which can seriously limit some real life data examples like image processing. Heller and Ghahramani (2007) used nonparametric data labeling using a Bayesian approach for overlapping clusters. They developed a infinite overlapping mixture model, for modeling overlapping clusters which allows data membership in unrestricted number clusters with unbounded clusters using

MCMC algorithms. Regression tree models using nonparametric clustering labeling tools have also been developed especially in image processing as described by Jancsary et al. (2012). It show the power of flexibility brought to the regression trees due to nonparametric labels.

In this work we are going to use nonparametric mixture model as described in chapter 4. In particular nonparametric mixture models using graph theory as described by Azzalini et al. (2014) will be considered and the outcome groups, here to be treated as labels/classes for the classification will be considered as analysed in section 4.6. Results were then be used as labels and hence classification can be administered.

5.3.1. Nonparametric Mixture Labels Evaluation

In order to assess the quality and evaluate the use of nonparametric mixture labels in data labeling, we will use nonparametricaly labeled data in a classification problem. This is achieved via a paired T test of the data on different classifiers at 5% level of significance using Weka 8.0 version. The classifiers considered included Naive Bayes, logistic regression, IBK, bagging, J48, and random forest. The classifier analysis was done on the labeled Cleveland heart disease data with identified groups used as labels for the class attribute. The original data by Detrano had six missing values for observations 88, 167, 193, 267, 288, and 303. Classification analysis was therefore, done twice in this work: first labeled data and with the missing values replaced using Mean Value method and secondly with all the missing value observations deleted. We will call the labeled data sets, *nclev* and *nclevNO* respectively, for the nonparametrically labeled data with replaced missing values and in which missing values where deleted. The analysis was done under the hypothesis that

Definition 5.3.1. ***Hypothesis***

$\mathbf{H_0}$: *There are no significant differences among the classifiers.*

$\mathbf{H_1}$: *There are significant differences among the classifiers*

Table 5.1. shows an output of results from a paired t test in which the Naive

Bayes was used as the test base. It shows that paired T test results using area under the Receiver Operating Characteristic (ROC) curve as the comparison field for the 6 classifiers.

Definition 5.3.2. ***Receiver Operating Characteristic (ROC)****-the positive sensitivity call/probability of the algorithm (True positive rate against False Positive rate).*

The ROC field was chosen in line with the discussion by [Demsär (2006)] that the area under ROC is a better test fields compared to other fields. The paired T tests where performed using experimenter in weka 8.0 version. The classifiers are defined as follows (1)-Naive Bayes, (2)-Logistic regression,

Table 5.1.: Naive Bayes as test base

Dataset	(1)	(2)	(3)	(4)	(5)	(6)
nclev	1.00	1.00	0.92 •	0.99	0.99	1.00
nclevNO	1.00	1.00	0.92 •	0.99	0.99	1.00

○, • statistically significant improvement or degradation

(3)- IBk-Linear neighbour distance search, (4)-Bagging, (5)- J48 and (6) is Random Forest. Results shows that IBk performs significant poorer than Naive Bayes whilst all the other classifiers performs significantly the same as Naive Bayes. Results using the IBk as the test base to check if there differences with the other classifiers are shown in table 5.2.. The results in

Table 5.2.: IBk as test base

Dataset	(3)	(1)	(2)	(4)	(5)	(6)
nclev	0.92	1.00 ○	1.00 ○	0.99 ○	0.99 ○	1.00 ○
nclevNO	0.92	1.00 ○	1.00 ○	0.99 ○	0.99 ○	1.00 ○

○, • statistically significant improvement or degradation

table 5.2. shows that the 5 classifiers are all significantly better that IBK for both data sets. Thus we deduce that when classifying heart disease patients using nonparametric model, we may use all the other 5 classifiers although IBk should be avoided as it significantly performs poorer.

5.3.2. Classifier Performance and Output

In this section we explored on every classifier mentioned here and summarised the performance measures for the two data sets. The measures considered here include percentage of correct classifications and Kappa statistics besides the ROC mentioned above.

Definition 5.3.3. ***Kappa Statistics****: a statistic which measures inter-rater agreement (degree of agreement among raters) for qualitative (categorical)*

Table 5.3. show the output results for the different classifiers for the two data sets. The output results shows that for both data sets the IBK

Table 5.3.: Nonparametric Labeling Classifier Results

Classifier	**nclev** % Correct	Kappa Statistic	**nclevNO** % Correct	Kappa Statistic
Naive Bayes	98.02%	0.96	98.99	0.97
Logistic regression	96.67%	0.99	98.65	0.97
Bagging	99.00%	0.98	98.99	0.98
IBK	93.07%	0.87	91.58	0.83
J48	99.00%	0.98	98.99	0.98
Random trees	99.00%	0.98	98.99	0.98

perform less accurately in both classifying observations and in the Kappa statistics as indicated by the highlighted blue values in table 5.3.. IBK is the least performer with correctly classified accuracy of 93.07% and 91.58 for nclev and nclevNO data sets respectively. The Kappa statistics for the same datasets are 0.86 and 0.83, respectively. This confirms the paired T test results from section 5.3.1, where the IBK performs significantly poorer than all the other classifiers and hence cannot be used for heart disease data. Again the performance from the nonparametrically labeled data where the missing values were imputed seems to perform better that removing the missing values.

5.4. Classification via Parametric Mixture Models

In this section, classification will be administered using same number of classifiers as in section 5.3.1 but using parametrically labeled data through Poisson mixture regression model as explained in chapter 3. The

labels were identified as high risk for class label 2 and low risk for class label 1. The detailed parametric mixture models are explained in chapter 3. This is a result established by [Mufudza and Erol (2016)] that heart disease prediction can be done categorically in 2 groups labeled as high risk and low risk groups. The individuals who were classified under category 1 are believed to have low heart disease risk whilst high risk individuals are classified as in group 2. The classifier or machine learning algorithms were administered on the labeled data with labels from a Poisson mixture model with two components.

5.4.1. Parametric Mixture Models Label Evaluation

The performance and accuracy of classifiers is here assessed using parametrically labeled data. The same hypothesis as defined in hypothesis 5.3.1 is followed, and a paired T test was performed under the hypothesis for 6 classifiers at 5% level of significance. Again, two data sets were used, **parclev** for replaced missing value data and **parclevNO** for data with deleted missing values. Table 5.4. show results of the paired T test using Naive Bayes as the base test classifier. It shows that when Naive Bayes is used as the test base, there are no significant differences with the IBK. However, there are significant differences were noted between Naive Bayes with logistic regression, bagging, J48 and random forest as they all perform significantly better than Naive Bayes. The result is true for both the **parclev** and **parclevNO** data sets. Using the IBK as a test base produces the same

Table 5.4.: Naive Bayes as test base

Dataset	(1)	(2)	(3)	(4)	(5)	(6)
parclev	0.91	0.98 ○	0.86	0.99 ○	0.96 ○	0.99 ○
parclevNO	0.91	0.99 ○	0.87	0.99 ○	0.96 ○	0.99 ○

○, • statistically significant improvement or degradation

Table 5.5.: IBK as test base

Dataset	(3)	(1)	(2)	(4)	(5)	(6)
parclev	0.86	0.91	0.98 ○	0.99 ○	0.96 ○	0.99 ○
parclevNO	0.87	0.91	0.99 ○	0.99 ○	0.96 ○	0.99 ○

○, • statistically significant improvement or degradation

result as indicated by table 5.5. for the same data sets. Results shows that there are no significant differences between the IBK and the Naive Bayes, whilst IBK significantly is different poor from logistic regression, bagging, J48 and random forest. Finally, table 5.6. represents the output of results

Table 5.6.: Random forest as base

Dataset	(6)	(1)	(2)	(3)	(4)	(5)
parclev	0.99	0.91 •	0.98	0.86 •	0.99	0.96 •
parclevNO	0.99	0.91 •	0.99	0.87 •	0.99	0.96 •

◦, • statistically significant improvement or degradation

using random forest as base. It indicates that there are no significant differences between random forest, logistic regression and bagging whilst Naive bayes, IBk and J48 performs significantly poorer than Random forest. Thus, we can conclude when parametric labeling is used, that Naive Bayes classifier and IBk are the least performers, followed by J48 and Logistic regression, bagging and random forest remained the highest performers in the classification. Thus we can use 3 classifiers when parametric labeling is used for heart disease data.

The classifier output for each of the classifier used for the six classifiers is given as table 5.7.. The tables shows output results for analysis done with both the parclev and parclevNO data sets being used for the Kappa statistics and percent correctly classified. Naive Bayes and IBK seems to

Table 5.7.: Parametric Labeling Classifier Results

Classifier	**parclev** % Correct	Kappa Stat.	**parclevNO** % Correct	Kappa Stat.
Naive Bayes	80.86%	0.58	81.81%	0.61
Logistic regression	**96.70**%	**0.93**	96.70%	0.93
Bagging	94.06%	0.87	94.27%	0.88
IBK	88.78%	0.75	88.21%	0.73
J48	95.05%	0.89	94.61%	0.88
Random trees	94.05%	0.87	95.28%	0.88

be the least performers for both data sets whilst logistic regression, bagging and random forest performs better. There seems to be better results with parclev than parclevNO although the difference is very slight. Thus we can actually conclude that deleting observations with missing values for the data

has a slightly negative effect on the performance of the classifiers.

5.5. Nonparametric versus Parametric labeling

In this section we are going to compare the classifier performance when both parametric and nonparametric labeling methods are used. A comparison analysis will be used to determine a better labeling technique. First paired T test analysis is done at 5% using Naive Bayes as the test base in weka experimenter and table 5.8. shows the output from the paired T test. The tables shows that when nonparametric labels are used in nclev

Table 5.8.: Naive bayes as test base

Dataset	(1)	(2)	(3)	(4)	(5)	(6)
nclev	1.00	1.00	0.92 •	0.99	0.99	1.00
nclevNO	1.00	1.00	0.92 •	0.99	0.99	1.00
parclev	0.91	0.98 ◦	0.86	0.99 ◦	0.96 ◦	0.99 ◦
parclevNO	0.91	0.99 ◦	0.87	0.99 ◦	0.96 ◦	0.99 ◦

◦, • statistically significant improvement or degradation

and nclevNO datasets, 4 other classifiers performs significantly high and the same as naive Bayes with the exception of IBK which performs significantly poorer. However, using parametric labeling in data sets parclev and parclevNO does not only seem to have a negative impact on the performance of all the classifiers since they have lower values, but naive Bayes performance is reduced to be the same as IBK. However, using the test base as random forest as shown in table 5.9. does not change the results on nonparametric labeled data as IBK, remains significantly inferior to all the other 5 classifiers which performs the same. Using parametric labeling however, results

Table 5.9.: Random forest as test base

Dataset	(6)	(1)	(2)	(3)	(4)	(5)
nclev	1.00	1.00	1.00	0.92 •	0.99	0.99
nclevNO	1.00	1.00	1.00	0.92 •	0.99	0.99
parclev	0.99	0.91 •	0.98	0.86 •	0.99	0.96 •
parclevNO	0.99	0.91 •	0.99	0.87 •	0.99	0.96 •

◦, • statistically significant improvement or degradation

in reduced performance for J48 as indicated in table 5.9. that Naive Bayes, IBK and J48 are significantly poor classifiers compared to random forest.

Logistic regression and Bagging however are not significantly different from random forest.

5.6. Summary

Therefore we can deduce that the use of nonparametric mixture models for data labeling has a greater impact on the performance of the classifier as it improves the performance. Analysis shows that when parametric mixture models are used for data modeling only 3 classifiers can be used which include, Logistic regression, Bagging and Random forest whilst the performance of Naive Bayes and J48 are reduced. On the other hand the use of nonparametric models improves the performances of the classifiers as shown by J48 and Naive Bayes performing equally the same as Random forest, bagging and logistic regression with the exception of IBK which remain the least performer in both instances.

Thus we recommend the use of nonparametric mixture models for data labeling compared to parametric mixtures.

6. RESULTS AND CONCLUSIONS

6.1. Introduction

The main focus of the study in this thesis was to develop and analyse model based clustering methods on heterogeneous populations. In particular with regard to the use of mixture models in both the parametric and nonparametric sense. This chapter gives a summary of the analysis, observations and recommendations on the use and application of mixture models highlighting both the parametric and nonparametric mixture models. Application to real data sets like the heart disease data were also done and recommendations on heart disease mixture models will be also be made thereof.

6.2. Use of Finite Mixture Models

Model based clustering was initially addressed and analysed via the use of finite mixture models. Finite mixture models were developed and analysed for mixed count data which had a lot of extra dispersion problems and abnormalities a sign of heterogeneity. Specifically the thesis outlines how Poisson Mixture models can be of help to deal with heterogeneous count data. Since generalised Poisson linear models applied to count data proved to be very poor models due unobserved heterogeneous features in the data (extra dispersion and non linearity challenges), use of GLM Poisson Mixture Models proved to fit the data better. This is because they cater for these dispersion challenges through different parameters in the mixture models in a componentwise format hence solving the extra dispersion challenges.

Finite mixture models uses the assumption that the data follows a mixture distribution formed of components which follows a Poisson distribution in this case but having same shape but different parameter values. In particular Zero inflated Poisson Mixture models seem to do well with both the data and dispersion challenges even in the presence of high frequency with high zero value. Thus we encourage the use of Poisson mixture models

on count data with unobserved heterogeneity and in particular the Zero Inflated Poisson Mixture models and its variations on data with high frequency of zeros.

6.2.1. Use of Finite Mixture Regression Models

The use of GLM mixture Poisson regression models was reviewed under different classes which includes standard and concomitant variable models. The general use of the mixture regression models was to capture the heterogeneity which cannot be catered for via a normal general Poisson regression models, by clustering homogeneous observations in components. This seemed to add a lot of value on the concomitant model than it does on the standard mixture model. Inclusion of mixture regression models greatly improves the model fit on the data as indicated by very low BIC values compared to GLM models. Thus improving on the efficiency at which we can predict heart disease since different components has different risk factors. They predict individuals component based on whether they level of risk involved. Mixture models also captures the extra variability via a concomitant variable Poisson mixture regression model. This tend to capture more heterogeneity than both the standard Poisson mixture regression model and well as the single component (General Poisson regression model not shown).

The presence of so many zeros within the count data also triggered us to incorporate a Zero inflated mixture Poisson regression model to handle the many zeros for Poisson models. The idea of handling the zeros in a special manner as expected improves the models since it does not only fit the data properly but also make realistic assumptions. This results in categorising those who are likely not to develop the disease into a single cluster whilst those with high chances of developing the disease are in another component. The assumption is that those with the disease are modeled by Poisson mixture regression to predict the diagnosis allowing variability on the covariates. We noted that the Zero inflated Poisson mixture model with a concomitant variable is even better than just a two Poisson mixture regression

model as it has better and lower BIC and AIC values. Thus mixture models gives more insight and better describes the data than supervised classification method does above its ability to efficiently predict heart disease and identify major risk factors.

6.3. Use of Nonparametric Mixture Models

Although finite mixture models performed well in capturing heterogeneity, the fact that the components had to follow a shape and density proved too rigid. Nonparametric mixture models were therefore introduced and analysed on count data so as to increase flexibility hence compact structures in the mixture model. It was noted however, that though flexible also come with computational challenges. In this work nonparametric mixture models where analysed in a two fold manner: The first one concentrated on the use of graph theory to determine the clusters and closely connected points via Voronoi graphs and paired edged connections whilst the second method used a conditional independence assumption on the components via EM like algorithm. In both cases the densities were then estimated by kernel density functions with a Gaussian kernel.

The use of density estimators with Gaussian kernels were explored for all the nonparametric mixture models, results show that whilst change in smoothing parameter/bandwidth has effects to mixture components, there were no effects with associated with change in shape as demonstrated in section 4.2.3 of chapter 4. Significant effects were however noticed with change in the bandwidth of the kernel both in figure 4.1. of section 4.2.3 and section 4.5 of chapter 4. In this regard both adaptive and fixed bandwidth were used in the model analysis and the adaptive performed better than fixed case. The adaptive bandwidth kernels deal with extra heterogeneity better as they are flexible per every observation that need to be clustered.

6.4. Data Labeling via Mixture Models

Mixture models were used in this thesis as a labeling technique to unlabeled data. A combination of both parametric and nonparametric mix-

ture models were analysed as data labeling tools. We analysed the labeled data via supervised classification through different classifier to evaluate the quality of the classes and hence better clustering method between the parametric and nonparametric mixtures. It was established that nonparametric mixtures models tend to produce better class labels if used for data labeling compared to parametric mixture models. This was evidenced by their both ability to improve classifier performance and very high area ROC and Kappa values compared to the parametric mixture model labeled models. We concluded the high quality labels/clusters maybe due the flexibility in dealing with heterogeneity in nonparametric mixture models as explained in chapter 4. In particular, for count data example, if effective and efficient methods for classifying heart disease patients has to be employed then we may have to employ nonparametric mixture models as a way of classifying patients over parametric mixture models.

6.5. Heart Disease Risks

We discovered that unlike previous prediction models on heart disease where the general risk factor were cholesterol, exercise and blood pressure without inferring into how much they affect. On the contrary when Poisson mixture regression models are used, it showed that males are at more risk than females for heart disease whilst cholesterol doesn't seem to be a dangerous risk factor on heart diseases. Chest pain is also indicated among the major risk factors with non anginal pain associated with heart disease more than angina pains, lack of exercise as before attributed to heart disease, those with abnormal resting electrocardiographic level are at higher risk, the type of defects on the chambers and number of major vessels an individual have. Therefore mixture models do not only fit the data better but also gives more information on the heterogeneous characteristics of different individuals in the prediction of heart diseases. We can also safely conclude that heart disease can be efficiently predicted by a two component Poisson mixture regression models, with each component describing attributes differently. Thus the

different components cater and captures the different heterogeneity properties exhibited by the general Poisson regression model hence the conclusion that model based clustering adds value to the prediction of heart diseases. The mean heart diagnosis is here given as a multiplicative factor of the parameter estimates and is differently predicted by two component concomitant Poisson mixture regression model, with the best model catering for the zeros in ZIP mixture regression model. The groups were the same as identified in Mufudza and Erol (2016) using Poisson Mixture regression models as low risk and high risk individual.

However, when nonparametric mixtures are used via graphical methods, a 2 component 2 cluster model best describes heart disease individuals using adaptive bandwidth for the kernel whilst fixed bandwidth favour a 2 component 3 cluster model. The adaptive bandwidth model proved a better model over the later. We can therefore deduce that the flexibility in the nonparametric mixture models can result in different number of clusters for the same number of components unlike the rigid assumption in the parametric case that clusters are represented as components. The conditionally independent assumption model for heart disease result in 2 component and 3 blocks model with the best compacted, well separated clusters.

In general therefore, heart disease diagnosis can be done using 2 categories : those under high risk and low risk individuals whether nonparametric or parametric mixture model are used as identified by Mufudza and Erol (2016). However, labels from nonparametric mixture models seems to be more ideal and in agreement they improve classifier perfomance under classification learning algorithms exposed in chapter 5. Thus we can conclude that flexibility on nonparametric mixture models brings the desired characteristics of unveiling heterogeneous properties and hence better labels for prediction models.

6.6. Further Research Recommendations

1. Finite mixture models included in this work used constant mixing proportions. We suggest the use of variable mixing proportions to the Poisson mixture regression models to bring in the extra flexibility required as suggested by Huang et al. (2013) a form of nonparametric mixture regression models. This can help bring stochasticity to the mixture model and hence heterogeneity can be compacted in a more holistic and practical way.
2. Components in the mixture models here used density functions of the same family, i.e Poisson distributions. A combination of varying mixing parameter and different density functions for different components can also be tried and evaluated.
3. Although nonparametric mixture models used here brought different levels of flexibility, by considering both EM like algorithms and graph theory model approach. clusters approximated from KDE with both adaptive and fixed bandwidths were considered. These also used Gaussian kernels were employed for the density estimates. Thus we recommend use of different density estimation methods other than Gaussian kernels to establish their effects on the number of components and clusters as well as the mixing proportions. These can also be substituted by a different way of estimating the densities like using smooth polynomials called smoothers.

BIBLIOGRAPHY

Akhil M., Chandra P., and Deekshatulu B.L, 2012. Heart Disease Prediction System using Associative Classification and Genetic Algorithm. ICECIT.

Anooj P., 2012. Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. Journal of King Saud University: Computer and Information Sciences, 24, 27-40.

Asparouhov T., Muthén B., 2015. Structural Equation Models and Mixture Models With Continuous Nonnormal Skewed Distributions. Structural Equation Modeling: A Multidisciplinary Journal, DOI: 10.1080/10705511.2014.947375.

Azzalin A., Menardi G., 2014. Clustering via Nonparametric Density Estimation: The R Package PdfCluster. Journal of Statistical Software, 57, 1-26.

Azzalini A.,Torelli N. 2007. Clustering via Nonparametric Density Estimation. Stat Comput, 17, 71-80. DOI 10.1007/s11222-006-9010-y

Bagirov A. M., Rubinov A. M., Soukhoroukova N. V., Yearwood Unsupervised and supervised data classification via nonsmooth and global optimization. School of Information Technology and Mathematical Sciences, The University of Ballarat. (unpublished).

Becker J. M., Rai A., Ringle C.M., *et al*, 2013. Discovering Unobserved Heterogeneity in Structural Equation Models to Avert Validity Threats. MIS Quarterly, 37, 665-694.

Benaglia T., Chauveau D., Hunter D., 2011. Bandwidth Selection in an EM-like algorithm for Nonparametric Multivariate Mixtures. World Scientific Publishing Co. Nonparametric Statistics and Mixture Models: A Festschrift in Honor of Thomas P. Hettmansperger, World Scientific Publishing Co., 15-27. hal-00353297.

Benaglia T., Chauveau D., Hunter R., 2009a. An EM-like Algorithm for semi- and Nonparametric Estimation in Multivariate Mixtures. Journal of Computational and Graphical Statistics, 18, 505-526.

Benaglia T., Chauveau D., Hunter R., Young D., 2009b. Mixtools: An R Package for Analysing Finite Mixture models. Journal of Statistical Software, 32, 1-29, http://www.jstatsoft.org/v32/i06/.

Bhatla N., Jyoti K., 2012. An Analysis of Heart Disease Prediction using Different Data Mining Techniques. IJERT, 1, 1-4.

Biau G., Cadre B., and Pelletier B., 2007. A Graph Based Estimator of the Number of Clusters. ESAIM: Probability and Statistics, 11, 272-280. DOI: 10.1051/ps:2007019.

Bockenholt U., 1993. The Analysis of "Pick any/n" Data With No prior Knowledge. Kwantitative Methoden, 42, 109-127.

Bonhomme S., Jochmans K., Robin J.M., 2016. Non-parametric estimation of finite mixtures from repeated measurements. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 78, 211-229, DOI: 10.1111/rssb.12110.

Boswell M.T., Ord J.K., Patil G.P., 1979. Statistical Distributions in Ecological Work. International Coperative Publishing House, USA, 4, 89-130.

Brochado A.O., Martins F.V., 2005. Assessing the Number of Components in Mixture Models. FEP, 194.

Chaurasia V., Pal S., 2013. Early Prediction of Heart Diseases Using Data Mining Techniques. Carib.j.SciTech, 1, 208-217.

Chauveau D., 2012. New mixture models and algorithms in the mixtools package.1 Seres Rencontres R, Bordeaux, France. <hal-00717545>.

Chauveau D., Hoang V.T.L., 2015. Nonparametric mixture models with Conditionally Independent Multivariate Component Densities. hal-01094837v2.

Chauveau D., Hoang V.T.L., 2016. Nonparametric Mixture Models with Conditionally Independent Multivariate Component Densities. Computational Statistics & Data Analysis, 103, 1-16. http://dx.doi.org/10.1016/j.csda.2016.04.013.

Cheng, Y.S. and Ray, S., 2014. Multivariate Modality Inference Using Gaussian Kernel. Open Journal of Statistics, 4, 419-434. http://dx.doi.org/10.4236/ojs.2014.45041.

Cuevas A., and Fraiman R., 1997. A Plug in Approach to Support Estimation. The Annals of Statistics, 25, 2300-2312.

Cuevas A., and Fraiman R., 2007. A Nonparametric approach to the estimation of Lengths and Surface Areas. The Annals of Statistics, 35, 1031-1051. DOI: 10.1214/009053606000001532.

Cuevas A., Febrero M., Fraiman R., 2000. Estimating the Number of Clusters. The Canadian Journal of Statistics, 28, 1-17.

Dai H., 2006. Parametric Estimation and Homogeneity Testing in Finite Mixture Models. The Graduate School, University of Kentucky.

Dalrymplea M.L; Hudsona I.L,. Ford, 2003. Finite Mixture, Zero-infated Poisson and Hurdle models with application to SIDS. Computational Statistics and Data Analysis, 41, 491-504.

Dangare C.S., Apte S.S., 2012. Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques. International Journal of Computer Applications, 47, 0975-888.

Dang U.K., Browne R.P., McNicholas P.D., 2015 Mixtures of Multivariate Power Exponential Distributions. The International Biometric Society, 1-11.

Demsar J., 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research, 7, 1-30.

Detrano R., M.D., Ph.D., Long Beach and Cleveland Clinic Foundation. V.A. Medical Center, *http://archive.ics.uci.edu/ml/datasets/Heart+Disease*.

Dillion W.R., Buckenholt U., et al., 1994. Issues in the Estimation and Application of Latent Structure Models of Choice. Marketing Letters 5, Kluwer Academic Publishers, 4, 323-334.

Erisöǵlu U., Erisoglu M., Erol H., 2011. A mixture of Two Different Distribution Approach to the Analysis of Heterogeneity in survival Data. World of Academy of Science, Eng. and Technology, 5, 06-29.

Everitt B.S., Hand D.J., 1981. Finite Mixture Distributions. Monography Applied Probability and Statistic, Chapman and Hall, 001: 10.1007/978-94-009-5897-5.

Fang R., 2008. Zero Inflated negative Binomial for Overdispersed Count Data with Excess Zeros and Repeated Measures, an Application to human Microbiota Sequential Data. NCCMM college.

Fraley C., Raftrey A.E., 1998. How many Clusters? Which clustering methods? Answers via Model Based Cluster Analysis. Computer Journal, 44, 578-588.

Fraley C., Raftery A.E., 2002. Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association, ABI/INFORM Global, 97, 611.

Fraley C., Raftery A.E., 2012. Mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. Technical Report No. 597 Department of Statistics, University of Washington, WA 98195-4322 USA.

Gillespie N.A., Neale M. C., 2006 A Finite Mixture Model for the Genotype and Environmental Interactions: Detecting Latent Population Heterogeneity. Twin Research and Human Genetics, 9, 412-423.

Goodman L.A., 2007. Explanatory Latent Structures Analysis using both Identifiable and unidentifiable Models. Boimetrika, 61, 215-231.

Gower, J.C., Ross, G.J.S., 1969. Minimum Spanning Trees and Single linkage Cluster Analysis. J.R.Stat. Soc., Series C (Applied Statistics), 18(1), 54-64.

Greene W., *Functional Form of Heterogeneity in Models for count Data*, Foundation and trends in economics, 1(2007), pp: 113-218.

Greene W., 2005. Functional Form and Heterogeneity in Models for Count Data. Foundations and Trends in Econometrics, 1, 113-218, DOI: 10.1561/0800000008.

Grün B., Leisch F., 2008. FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters.Journal of Statistical Software, 28(4), 1-35, doi:10.18637/jss.v028.i04

Grün B., and Leisch F., 2007. Fitting finite mixtures of generalized linear regressions in R. Computational Statistics & Data Analysis, 51(11), 5247-5252,doi:10.1016/j.csda.2006.08.014

Hair J.F(Jnr)., Anderson R.E., Tatham R.L., Black W.C., 1998. Multivariate Data Analysis. Prentice Hall International, 5th edition.

Hall P., Zhou X., 2003. Nonparametric Estimation of Component Distributions in a Multivariate Mixture. The Annals of Statistics, 31, 201-224.

Erol H., 2014. Mixture of Mixed distribution Models. 9th International Statistics days Symposium, Antalya, Turkey, 1-21. (unpublished)

Härdle W. 1990. Applied Nonparametric Regression. Cambridge University Press.

Härdle, W., Müller, M., Sperlich S., Werwatz A., 2004. Nonparametric and Semiparametric Models. Springer.

Hartigan J.A., 1975. Clustering algorithms. Wile Series in Probability and Mathematical Statistics.

Hayfield T., Raccine J.S., 2008. Nonparametric Econometrics: The np Package. Journal of Statistical Software, 27, 1-32.

Heller, K.A., Ghahramani, Z., 2007. A Nonparametric Bayesian Approach to Modeling Overlapping Clusters. 11th International Conference on AI and Statistics (AISTATS 2007).

Hilbe J. M., 2014. Modeling Count Data. Cambridge University Press.

Hossein R., Mash'al M., 2015. An Estimation Toolbox for Mixture Models. Journal of Machine Learning Research, 1, 1-5.

Huang M., Yao W., 2012. Mixture of Regression Models With Varying Mixing Proportions: A Semiparametric. J Am Stat Assoc., 107(498), 711-724, DOI: 10.1080/01621459.2012.68254.

Huang M., Runze L., Shaoli W., 2013. Nonparametric Mixture of Regression Models. J Am Stat Assoc, 108, 929-941.

Hunter D.R., Wang S., Hettmansperger T.P., 2007. Inference for Mixtures of Symmetric Distributions. The Annals of Statistics, 35, 224-251. DOI: 10.1214/009053606000001118.

Hunter D.R., Young D.S., 2012. Semiparametric mixtures of regressions. Journal of Nonparametric Statistics, 24(1), 19-38, DOI: 10.1080/10485252.2011.608430.

Jancsary, J., Nowozin, S., Sharp, T., and Rother C., 2012. Regression Tree FieldsŮ An Efficient, Non-parametric Approach to Image Labeling Problems. 00, 2376-2383, doi:10.1109/CVPR.2012.6247950.

Karlisa D., Meligkotsidoub L., 2007. Finite mixtures of multivariate Poisson distributions with application. Journal of Statistical Planning and Inference, 137, 1942-1960. www.elsevier.com/locate/jspi

Kasahara, H. and Shimotsu, K. 2010. Nonparametric Identification of Multivariate Mixtures. Technical report.

Kaur B., Singh W., 2014. Review on Heart Disease Prediction System using Data Mining Techniques. IJRITCC, 2, 3003-3008. ISSN:2321-8169.

Kedem B., Chiu, L. S., 1990. Estimation of Mean rain rate: Application to Satellite observations. Journal of Geographical research , 95, 1965-1972.

Kessler D., McDowell A., 2012 Introducing the FMM Procedure for the Finite Mixture Model. SAS institute, Cary.

Levine M., Hunter D., Chauveau B., 2011. Maximum Smoothed Likelihood for Multivariate Mixtures. Biometrika, Oxford University Press (OUP), 98, 403-416. <10.1093/biomet/asq079>. <hal-00516391>.

Liesenfield R., 1998. Dynamic Bivariate Mixture Models: Modeling the Behavior of Prices and Trading. Journal of Business & Economic Statistics, 16, 101-109.

Li J., Ray S., Lindsay B.G., 2007. A Nonparametric Statistical Approach to Clustering via Mode Identification. Journal of Machine Learning Research, 8, 1687-1723

Li J., Zha H., 2006. Two-way Poisson Mixture models for Simultaneous document classification and word clustering. Computational Statistics and Data Analysis, 50(1), 163-180. MR2196228.

Linden W.J., 1997. A Lognormal Model for Response Times on Test Items. Journal of Educational and Behavioral Statistics, 31, 181-204.

Lindsay B. G., Lesperance M. L., 1995. A review of semiparametric mixture models, Journal of Statistical Planning and Inference, 47 , pp: 29-39.

Liu Y., Li Z., Xiong H., Gao X., Wu J., 2010. Understanding of Internal Clustering Validation Measures. IEEE International Conference on Data Mining, 911-916.

López, M.I., Romero L.C., and Venture S., 2012. Classification via clustering for predicting final marks based on student participation in forums. 148-151.

Lubke G.H., Muthen B., 2005. Investigating Heterogeneity with Factor Mixture Models. Physiological Methods, 10, 21-39.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., 2016. cluster: Cluster Analysis Basics and Extensions. R package version 2.0.5.

Magidson J., Vermut J.K., 2002. A nontechnical Introduction to Latent Models. Statistical innovation Inc.

Magidson J., Vermut J.K., 2002b. Latent Class Cluster Analysis. Applied latent Class Analysis, Cambridge University Press, 89-106.

Magidson J., Vermut J.K., 2003. Latent Class Models for Classification. Computational Statistics & Data Analysis, 41, 531-537.

Mallapragada P.K.,Jin R., Jain A., 2010. Non-parametric Mixture Models for Clustering. Chapter: Structural, Syntactic, and Statistical Pattern Recognition, 6218 of the series Lecture Notes in Computer Science, 334-343.

Masethe H.D., Masethe M.A., 2014. Prediction of Heart Disease using Classification Algorithms. WCECS-USA, ISSN: 2078-0966.

Ma Y., Yao W., 2012. Flexible estimation of a semiparametric two component mixture model with one parametric component. Electronic Journal of Statistics, 9, 444-474, DOI: 10.1214/15-EJS1008.

McLachlan G.J., Krishnan T., 1997. The EM Algorithm and extension. John Wileys and Sons-Applied probability and statistics.

McLachlan G.J., Basford K.E., 1988. Mixture models: inference and applications to clustering. Marcel Dekker, United States, 84, ISBN 0824776917.

McLachlarin G.J., Peel D., Basford K.E., Adams P., 1999. The EMMIX Algorithm for the Fitting of Normal and t-Components. Jstatsoft.org.

McLachlarin G., Peel D., 2000. Finite Mixture Models. John Wiley and Sons series, Applied probability and Statistics.

McLachlan G.J., Rathnayake S., 2014. On the Number of Components in a Gaussian Mixture model. WIREs Data mining Knowl Discov. 10, 1002-1135.

McNicholas P.D., 2011. On Model-Based Clustering, Classification, and Discriminant Analysis. JIRSS , 10, 181-199.

McNicholas P.D., Murphy T.B., 2010. Model-Based Clustering of Microarray Expression Data via Latent Gaussian Mixture Models. Bioinformatics, 26, 2705-2712.

McParland, D., Gormley, I.C., 2016. Model based clustering for mixed data: clustMD. Adv Data Anal Classif, 10, 155. doi:10.1007/s11634-016-0238-x.

Menardi G., 2011. Density-based Silhouette diagnostics for clustering methods. Stat Comput, 21, 295-308. DOI 10.1007/s11222-010-9169-0.

Menardi G., Azzalini A., 2014. An Advancement in Clustering via Nonparametric Density Estimation. Stat Comput, 24, 753-767. DOI 10.1007/s11222-013-9400-x.

Mozzaffarian D., and others., 2015. American Heart Statistics Committee and Stroke Statistics subcommittee. Heart disease and stroke Statistics, doi:10 1161/CIR.0000000000000152.

Mufudza C., Erol H., 2016 Poisson Mixture Regression Models for Heart Disease Prediction. Computational and Mathematical Methods in Medicine, Article ID 4083089, 10 pages http://dx.doi.org/10.1155/2016/4083089.

Muthen B. O., 1989. Latent Variable Modeling In Heterogeneous Populations. PSYCHOMETRIKA, 54, 557-585.

Newcomb S., 1986. A Generalized Theory of the Combination of observations so as to obtain the Best Result. American Journal of Mathematics, 8, 343-366.

Newman M.E.J., *et al*., 2007. Mixture Models and Exploratory Analysis in Networks. Proc Natl Acard Sci., DOI:10.1073/pnas.0610537104.

Ng S.K., McLachlan G.J., McGiffin D.C., O'brien M., 1999. Constrained Mixture Models in Competing Risks Problems. Environmetrics, 10, 753-767.

Ng Shu-kay, Mclachlan G.J., Wang K., Jones L.B., Ng S. W., 2006. A Mixture model with random-effects components for clustering correlated gene-expression profiles. Bioinformatics/computer Applications in The Biosciences - BIOINFORMATICS, 22, 1745-1752.

Pearson K., 1894. Contribution to the mathematical theory of evolution. Philosophical Transactions of the Royal Society, 185, 71-110.

Raccine J., Li Q., 2004. Nonparametric Estimation of Regression Functions with Both Categorical and Continuous Data. Journal of Econometrics, 119, 99-130.

Raftrey A.E., Banfield J.D., 1993. Model Based Gaussian and Non-Gaussian Clustering. Biometrics, 54, 803-821.

Rebafka T., Roueff F., 2015. Nonparametric estimation of the mixing density using polynomials. Mathematical Methods of Statistics, Allerton Press, Springer, 3, <hal-00458648v2>.

Rinaldo A., Wasserman L., 2010. Generalised Density Clustering. The Annals of Statistics, 38, 2678-2722. DOI: 10.1214/10-AOS797.

Rinaldo A., Singh A., Nugent L., 2012. Stability of Density Based Clustering. Journal of Machine Learning Research, 13, 905-948.

Romero L.C., López, M.I., Luna J.M., and Ventura S. 2013. Predicting students final Performance from Participation in On-line Discussion Forums. Computers & Education, 68, 458-472.

Ruggeri F., Soriano J., 2014. Semiparametric Bayesian models for clustering and classification in the presence of unbalanced in-hospital survival. Appl. Statist., 63, 25-46.

Sakai T., du Plessis M., Niu G., Sugiy M., (2016) Semi-Supervised Classification Based on Classification from Positive and Unlabeled Data. arXiv:1605.06955v3.

Schlattmann P., 2009. Medical Application of Finite mixture Models, Springer, 1, ISB. 978-3-540-68650-7.

Scott D.W., Sain S., 2005. Multi-dimensional Density Estimation. Hand Book of Statistics, 24, 29-261.

Sgouritsa E., Janzing D., Peters J., Scholkopf B., 2013. Identifying Finite Mixtures of Nonparametric Product Distributions and Causal Inference of Confounders.

Shimizu K., 1993. A Bi-variate Mixed Log-normal distribution with Analysis in rainfall Data. American Meteorological Society, 32, 161-171.

Smyth P., 1999. Probabilistic Model Based Clustering of Multivariate and Sequential Data. University of California.

Soni J., Ansani U., Sharma D., Soni S., 2011. Predictive Data Mining for Medical Diagnosis?: An Overview of Heart Disease Prediction. IJCA, 17, 43-48.

Stuetzle, I., Nugent, R. 2010. A Generalised Single Linkage Method for Estimating the Cluster Tree of a Density. J. Comput. Graph Statistics, 19, 397-418.

Suhaila J., Ching-Lee K., *et al*., 2011. Introduction to Mixed Distribution in Fitting Rainfall Data. Open Journal of Modern Hydrology, 1, 11-22.

Suhaila J., Yee K., Yusof F., 2013. Effect of Zero Measurement in Rainfall data. Jurnal Teknologic, 63, 35-39.

Sultan K.S., Ismail M.A., Al-Moisheer A.S., 2007. Mixture of Two Inverse Weibull Distributions: Properties and Estimation. Computational Statistics & Data Analysis, 51, 5377-5387. www.elsevier.com/locate/csda

Sung H.G., 2004. Gaussian Mixture Regression and Classification. Rice University, Doctor of Philosophy Thesis, 1-29.

Taneja A., 2013. Heart Disease Prediction System Using Data Mining Techniques. Oriental Journal of Computer Science and Technology, 6, 457-466.

Tang Y., Browne R.P., McNicholas P.D., 2015. Model based clustering of high-dimensional binary data. Computational Statistics and Data Analysis, 87, 84-101.

Vermut J. K., 2005. Introduction to Latent Class and Finite Mixture Modelling. Department of Methodology and Statistics, Tilburg University, The Netherlands, 1-30.

Vichi M., 2008. Fitting Semi-parametric Clustering Models to Dissimilarity Data. Adv Data Anal Classif, 2, 121-161.DOI 10.1007/s11634-008-0025-4.

Wang K., Yau K.K.W., Lee A., and Mclachlan G.J., 2007. Two-Component Poisson Mixture Regression Modeling of Count Data with Bivariate Random Effects. Mathematical and Computer Modeling, 46, 168-1476.

Wedel M., 2002. Concomitant Variables in Finite Mixture Models. Statistica Neerlandica, 56, 362-375.

Wei Yů, McNicholas P.D., 2015. Mixture model averaging for clustering. Adv Data Anal Classif, 9, 197-217, DOI 10.1007/s11634-014-0182-6.

Witten D. M., Tibshirani R., 2011. Penalized classification using FisherŠs linear discriminant. J. R. Statist. Soc., 73(5), 753-772. 10.1111/j.1467-9868.2011.00783.x

Xiang S., Yao W., 2016. Mixture regression with Single Index.1-29, arXiv:1602.06610v1 [stat.ME].

Young D., Benaglia T., {et al }, 2016. Tools for Analyzing Finite Mixture Models. R:Package mixtools, https://cran.r-project.org/web/packages/mixtools/mixtools.pdf.

Zhang H., Huang Y., 2015. Finite Mixture Models and their applications. Austin Biometrics and Biostatistics, 2, id1013

Zhang W., Fan J., Sun Y., 2009. Semiparametric Model for Cluster Data. The Annals of Statistics, 37, 2377-2408, DOI: 10.1214/08-AOS662.

Zhou M., Carin L., 2013. Nagative Binomial Process Count and Mixture Modeling. IEEE.

Zhu X., Hunter D.R., 2015. Clustering Via Finite Nonparametric ICA Mixture Models. Annals of Applied Statistics, 1-27.

AUTHOR'S BIBLIOGRAPHY

Chipo Mufudza was born in 1981 in Buhera Zimbabwe as a second child in a family of five. She did her primary and high school education at Ndongwe primary school and Nyashanu High school in Buhera, respectively. She was then admitted for a 4 year BSc(hons) in Applied Mathematics at the National University of Science and Technology (NUST) in Zimbabwe between 2001-2005. Having graduated with a second best class she got a scholarship under NUFU to study her Master of Science degree in Mathematics at the University in 2005-2007 at the University of Zimbabwe in Zimbabwe. This was the beginning of her involvement in bio-mathematical research with health applications. She has publications in HIV, Cancer, Tuberculosis, heart disease to date using both qualitative and quantitative techniques.

In 2012 she got a Turkish government scholarship to pursue a PhD degree in Statistics at Çukurova University in Turkey which she has recently finished. It is through this time that she was also able to learn Turkish language to a C1 level to add to her 3 other languages including English. Over the years her research interests have also been shaped to areas of data mining in big data, model based clustering and mixture models. She also has publications in the area of mixture models using and data mining which are health centered. Have a desire to continue in solving biomedical challenges through data science.

APPENDIX

APPENDIX A

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2016, Article ID 4083089, 10 pages
http://dx.doi.org/10.1155/2016/4083089

Research Article

Poisson Mixture Regression Models for Heart Disease Prediction

Chipo Mufudza and Hamza Erol

Statistics Department, Cukurova University, 01330 Adana, Turkey

Correspondence should be addressed to Chipo Mufudza; chipmuf@gmail.com

Received 20 April 2016; Accepted 20 October 2016

Academic Editor: David A. Winkler

Copyright © 2016 C. Mufudza and H. Erol. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Early heart disease control can be achieved by high disease prediction and diagnosis efficiency. This paper focuses on the use of model based clustering techniques to predict and diagnose heart disease via Poisson mixture regression models. Analysis and application of Poisson mixture regression models is here addressed under two different classes: standard and concomitant variable mixture regression models. Results show that a two-component concomitant variable Poisson mixture regression model predicts heart disease better than both the standard Poisson mixture regression model and the ordinary general linear Poisson regression model due to its low Bayesian Information Criteria value. Furthermore, a Zero Inflated Poisson Mixture Regression model turned out to be the best model for heart prediction over all models as it both clusters individuals into high or low risk category and predicts rate to heart disease componentwise given clusters available. It is deduced that heart disease prediction can be effectively done by identifying the major risks componentwise using Poisson mixture regression model.

1. Introduction

Heart disease encloses a number of conditions that influence the heart and not just heart attacks. These may include functional problems of the heart such as heart-valve abnormalities, high blood pressure (BP), smoking, diet, cholesterol, or irregular heart rhythms. Problems like these can lead to heart failure, arrhythmia, and a host of other problems. The work in [1] claims that heart diseases have become the leading global death accounting for 17.3 million deaths per year and killing 1 in 7 in the US alone. Therefore effective and efficient automated heart disease prediction systems can be beneficial to both the patient and cardiologist. Although there has been increasing interests on heart disease problems especially with the use of data mining techniques and algorithms, most of them concentrated on a supervised classification approach through different classifications and algorithms.

In a comparative approach research by Bagirov et al. [2] they showed that it is possible to classify heart disease problems using either supervised or unsupervised classifications. Supervised classification on different patients status by data mining algorithms to predict heart disease has been explored by various authors in machine learning. H. D. Masethe and M. A. Masethe [3] identified that the predictive accuracy of J48, over REPTREE and CART, was reliable for heart disease prediction in South Africa. Weighted fuzzy models based on supporting have also been studied and analysed showing an improvement against the network based models [4]. Supervised classification algorithms can improve the efficiency to cardiologist as shown by Taneja [5] on PGI data where over 7300 observations were classified using J48 and Naive Bayes in WEKA. In a general review on the heart disease using data mining techniques done by Kaur and Singh [6], they summarised that most researches show that the main risk factors are cholesterol, lack of exercise, obesity, and high blood pressure whilst the best algorithms seem to be dominated by decision trees.

Supervised classifications predictions can also be improved in some cases by incorporating unsupervised classification techniques like clustering as a preprocessing procedure. However, it may not always be the case as shown by Soni et al. [7] that decision trees can still outperform the Bayesian classification even when clustering is incorporated although they mentioned that both algorithms can be improved by genetic algorithms. They also implemented a combination of associative classification and genetic algorithms in an effort to come up with the best system

Heart Disease diagnosis via Nonparametric Mixture Models

Chipo MUFUDZA[1] and Hamza EROL[1]

[1]*Department of Statistics, Çukurova University, 01330, Adana, Turkey*

chipmuf@gmail.com
herol@cu.edu.tr

ABSTRACT

Effective and efficient ways to heart disease diagnosis can be improved via clustering individuals with heterogeneous characteristics to similar risk groups. This paper focuses on nonparametric density based cluster analysis on the risks of heart disease via nonparametric mixtures. Cluster density distributions for the nonparametric mixture model are done through Gaussian kernel density estimators using graph theory techniques. The cluster quality for the clusters from the models were analysed and diagnosed via a density based silhouette information criteria. Although the number of components is not assumed the same with clusters, results shows that individuals under heart disease risks can be grouped into two categories using two component model. It was also concluded that the individuals in different cluster have varying risk levels for heart disease.

ACKNOWLEDGEMENT

This research was made possible due to the support form Turkish government Scholarship. Thank you very much.

REFERENCES

[1] C. Mufudza, H. Erol, Computational and Mathematical Methods in Medicine, Article ID 4083089, 10 pages (2016).
[2] G. Mernadi, Stat Comput., 21, 295–308 (2011).
[3] J. Li, S. Ray, B.G. Lindsay, J. Machine Learning Research, 8, 1687-1723 (2007).
[4] S. Bonhomme, K. Jochmans, J.M. Robin, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 78, 211-229 (2016).
[5] A. Azzalini, N. Torelli, Stat Comput., 17, 71-80 (2007).
[6] A. Azzalini, G. Menardi, Journal of Statistical Software, 57, 1-26 (2014).
[7] G. Menardi, A. Azzalini, Stat. Comput., 24, 753-767 (2014).

APPENDIX B

This section concentrate on exploring the data set we are going to use for all the examples in the thesis as given by Detrano

Table 0.1.: Cleveland Heart Disease Variables

NAME	TYPE	DESCRIPTION
Age	Continuous	Age in years
Sex	Discrete	$1 = male, 0 = female$
Cp	Discrete	Chest pain type: 1 typical angina 2: atypical angina 3: non angina pa 4: asymptotic
Trestbp	Continuous	Resting blood pressure(mm/Hg)
Chol	Continuous	Serum cholesterol (mg/dl)
Fbs	Discrete	Fasting blood sugar $> 120 mg/dl$: $1 = true$ $0 = false$
Restecg	Discrete	Resting electrocardiographic results: $0 =$ normal $1 =$ having ST-T wave abnormality $2 =$ showing probable or define left ventricular hypertrophy by Estes?riteria
Thalcd	Continuous	Maximum heart rate achieved
Exang	Discrete	Exercise induced angina: $1 = yes$ $0 = no$
Oldpeak	continuous	ST depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment : $1 = upsloping$ $2 = flat$ $3 = downsloping$
Ca	discrete	number of major vessels (0-3) colored by flourosop
Thal	discrete	3 = normal; 6 = fixed defect; 7 = reversable defect
num	dicsrete	diagnosis of heart disease (angiographic disease status) Value $0 :< 50\%$ diameter narrowing Value $1 :> 50\%$ diameter narrowing

APPENDIX C

The distribution density functions for all the variables were estimated by kernel density estimators and these estimated densities shows different levels and properties of each of the variables which are the risks involved in heart disease diagnosis. This was done for both continuous or categorical variablse as shown by results in figures (0.1., 0.2., 0.3.): Figure

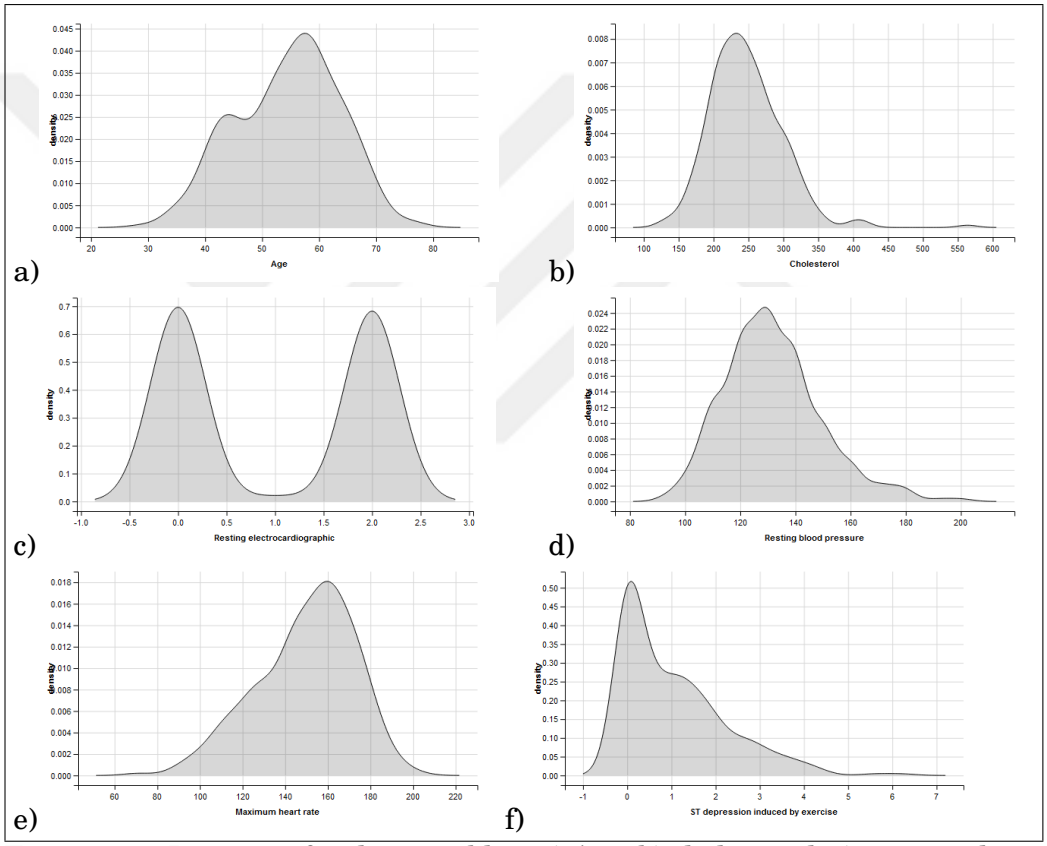


Figure 0.1.: Densities for the variables: a) Age, b) cholesterol, c) resting electrocardiographic, d) resting blood pressure, e) Maximum heart rate, f) ST depression

(0.1.) shows that although age is a continuous variable it has 2 peaks and is negatively skewed as shown by figure (0.1.a) an indication of heterogeneity within the age variable indicating about 2 groups. Cholesterol shows a slightly almost normally distributed with slight positive skewness an indica-

tion of possible existence of outliers with the variable as indicated by figure (0.1.b). Figure (0.1.c) shows the density of the electrocardiographic with 2 groups as expected since it has two levels. Resting blood pressure seem to be high from low to medium values than very high values as indicated by the positive skewness of figure (0.1.d) with the small high figure giving existence of the second group. Maximum heart rate seem to be highly negatively skewed as shown by figure (0.1.e) whilst St depression is highly positively skewed as in figure (0.1.f) with both variable hinting 2 categories of people.

The density distribution estimates for sex, chest pain, fasting blood sugar,

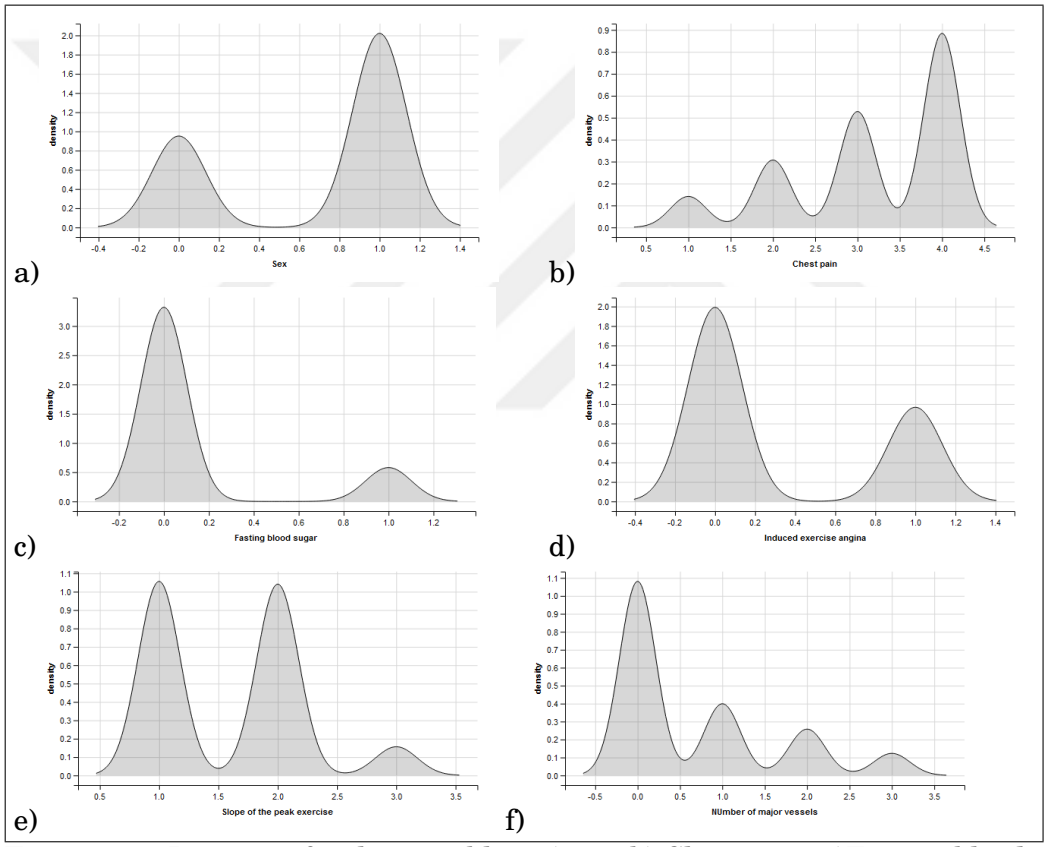
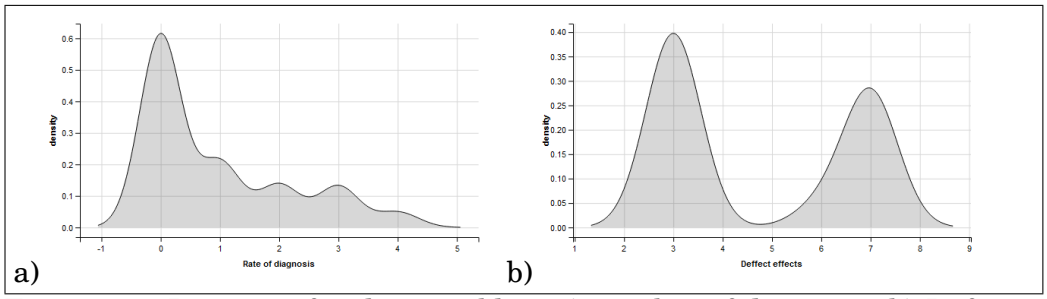


Figure 0.2.: Densities for the variables: a) sex, b) Chest pain, c)Fasting blood sugar, d) Exercise induced angina, e) slope of peak exercise, f) Number of major vessels

exercise induce angina, slope of peak exercise and number of major vessels are shown by figure (0.2.a, b, c, d, e, f) respectively. They all sow the groupings with each group as expected from the levels given in each of the variable.

The number of diagnosis density function shows 5 peaks an indication of the



a) b)

Figure 0.3.: Densities for the variables: a) number of diagnosis, b) Defects effects

levels of the variable as given by figure (0.3.a) whilst figure (0.1.b) shows the density estimate function for the defects effects.

APPENDIX D

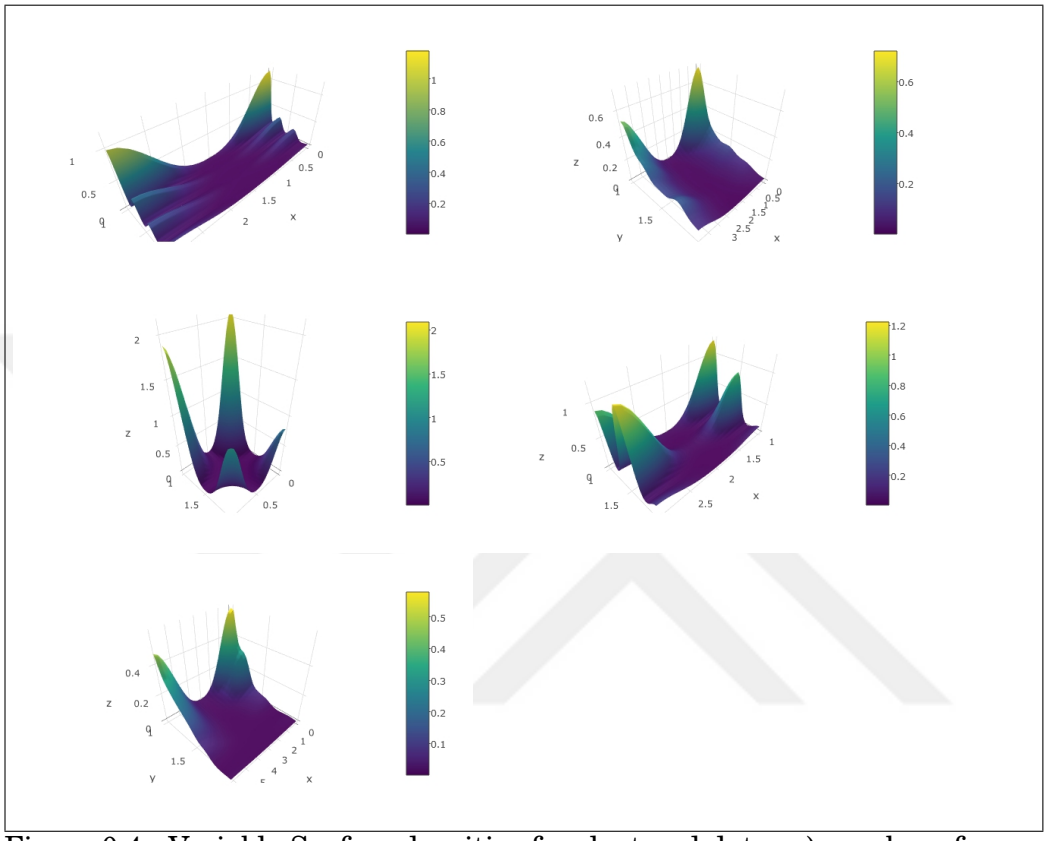


Figure 0.4.: Variable Surface densities for clustered data: a) number of vessels, b) Number of diagnosis , c) exercise induced angina d) slope of exercise and e) depression induced by exercise

APPENDIX E

Block 1:Coordinates 1 4 5 8 10

	Mean	52.6	130	233	156	0.821
	std	1	4	5	8	10
C1	1	82.10	47.60	47.10	-93.30	1.98
	4	47.60	254.00	82.30	-32.70	4.36
	5	47.10	82.30	1780.00	146.00	-6.45
	8	-93.30	-32.70	146.00	440.00	-7.02
	10	1.98	4.36	-6.45	-7.02	1.12

	Mean :	57.3	134	269	140	1.39
	Std	1	4	5	8	10
C2	1	66.600	31.10	71.30	-14.80	0.718
	4	31.100	388.00	96.30	39.70	1.820
	5	71.300	96.30	3260.00	122.00	4.670
	8	-14.800	39.70	122.00	490.00	-6.710
	10	0.718	1.82	4.67	-6.71	1.500

Block 2: Coordinates 2 6 9 14

	Mean	0.695	0.163	0.566	1.58
	Std	2	6	9	14
C1	2	0.2120	-0.01890	0.07020	0.1900
	6	-0.0189	0.13600	-0.00638	0.0344
	9	0.0702	-0.00638	0.24600	0.1370
	14	0.1900	0.03440	0.13700	1.5400

	Mean	0.695	0.163	0.566	1.58
	Std	2	6	9	14
C2	2	0.2120	-0.01890	0.07020	0.1900
	6	-0.0189	0.13600	-0.00638	0.0344
	9	0.0702	-0.00638	0.24600	0.1370
	14	0.1900	0.03440	0.13700	1.5400

Block 3 Coordinates 3 7 11 12 13

	Mean	2.89	0.82	1.53	0.392	4.27
	Std	3	7	11	12	13
C1	3	0.9070	-0.0283	0.0228	0.0217	0.2410
	7	-0.0283	0.9570	0.0867	0.0169	0.0675
	11	0.0228	0.0867	0.4100	0.0133	0.2920
	12	0.0217	0.0169	0.0133	0.6670	0.3520
	13	0.2410	0.0675	0.2920	0.3520	3.2700

	Mean:	3.59	1.26	1.72	1.17	5.45
	Std	3	7	11	12	13
C2	3	0.6370	0.0222	0.1150	0.0458	-0.3100
	7	0.1150	0.0225	0.3060	0.0468	0.2820
	11	0.1630	0.0458	0.0468	0.8690	0.0873
	12	0.3670	-0.3100	0.2820	0.0873	3.6400
	13	0.2410	0.0675	0.2920	0.3520	3.2700